Electronic Journal of Statistics
ISSN: 1935-7524

Moment-Based Adjustments of Statistical Inference in High-Dimensional Generalized Linear Models

Kazuma Sawaya[1], **Yoshimasa Uematsu**[2] **and Masaaki Imaizumi**[3]

[1] *The University of Tokyo, e-mail:* sawaya@g.ecc.u-tokyo.ac.jp

[2] *Hitotsubashi University, e-mail:* yoshimasa.uematsu@r.hit-u.ac.jp

[3] *The University of Tokyo, e-mail:* imaizumi@g.ecc.u-tokyo.ac.jp

Abstract: We develop a statistical inference method for generalized linear models (GLMs) in high-dimensional settings, where the number of unknown coefficients p is of the same order as the sample size n. In this regime, constructing confidence intervals requires estimating unknown hyperparameters, such as the signal strength. However, existing estimators for the hyperparameters are not stably applicable to GLMs when p/n is close to or greater than 1, both theoretically and empirically. In this study, we develop an estimator for the hyperparameter that addresses the issue and establish an inferential framework, provided that the link function of the GLM exhibits an asymmetry property. The proposed estimator utilizes the moments of the output variable of GLMs and a convex surrogate loss. Our framework is theoretically valid even when the limit of p/n exceeds 1, ensuring the strong consistency of the hyperparameter estimator and asymptotically attaining the exact coverage probability of the confidence intervals. Our numerical experiments support these theoretical results.

MSC2020 subject classifications: Primary 62F99, 62J12; secondary 62G05.
Keywords and phrases: bias correction, confidence interval, state evolution, proportionally high-dimension.

1. Introduction

We consider a pair $(\boldsymbol{X}, Y)$ with p-dimensional random feature $\boldsymbol{X}$ and random response Y following the generalized linear model (GLM):

$$\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] = g\left(\boldsymbol{x}^\top \boldsymbol{\beta}\right), \quad \forall \boldsymbol{x} \in \mathbb{R}^p, \tag{1}$$

where $g : \mathbb{R} \to \mathbb{R}$ is an inverse link function that monotonically increases, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^p$ is an unknown deterministic coefficient vector. This is one of the most popular classes of statistical models, including linear regression model, logistic regression model, Poisson regression model, and others. Given a sample of size n and assuming n-dependent dimension $p \equiv p(n)$, a GLM and its specific models in the *proportionally high-dimensional regime*, where there exists $\kappa > 0$ such that

$$n, p(n) \to \infty \quad \text{and} \quad p(n)/n \to \kappa \in (0, \infty), \tag{2}$$

have recently been studied [BM11, Ran11, EKBB$^+$13, TAH18, SC19, BKM$^+$19, Bel22].

Our interest lies in *statistical inference* (or simply *inference*), such as hypothesis testing and constructing confidence intervals, for $\boldsymbol{\beta}$. In high-dimensional settings (2), a maximum likelihood estimator (MLE) is influenced by dimensionality κ and unknown parameters in a complex manner [SCC19, SC19, SAH19, BKM$^+$19, FVRS22], which is contrast to low-dimensional settings [McC80, Cor83] or other (sparse) high-dimensional settings [VdGBRD14, TT17, FL21, CGM23, PSVAV24]. Indeed, the asymptotic bias and variance can be characterized as fixed points of a nonlinear system called the *state evolution (SE) system*. The several unknown hyperparameters in the system must be estimated for inference, which is a major challenge for practical use.

Despite the development of such inferential frameworks, some more challenges remain. First, estimating the SE hyperparameters becomes hard as the dimensionality increases, particularly when $\kappa \geq 1$. Existing estimators, including [SC19, YYMD21, CLM24], have been theoretically justified only for cases where $\kappa < 1$. Moreover, some methods exhibit numerical instability even when κ is less than 1 but close to 1, as we verify in Section 5. Second, even if the SE hyperparameters were known, the inference method with a high-dimensional MLE proposed by [SC19] would not be applicable to certain GLM settings. This is because, in some GLMs, the negative log-likelihood function is non-convex, rendering existing methods relying on convexity inapplicable.

In this study, we develop a *moment-based adjustment* for statistical inference in high-dimensional GLMs and prove its asymptotic validity. This methodology is based on the following techniques. First and most importantly, we construct a simple yet effective estimator of the SE hyperparameter, which can be computed from the data by leveraging a moment equation on the output variable Y. This method enables the estimation of the SE hyperparameters regardless of the value of $\kappa > 0$, while requiring the link function $g(\cdot)$ to be *asymmetric*, i.e., $g_0(x) := (g(x)+g(-x))/2$ being strictly monotone. The moment-based estimator remains applicable even when $\kappa \geq 1$, as it neither depends on the inverse of sample covariance matrices nor on the estimator of $\boldsymbol{\beta}$. The second technique is the use of *surrogate* loss function, which is a convex and consistent loss function for GLMs. This loss function mitigates the non-convexity of the negative log-likelihood function in a certain class of GLMs. We leverage this loss and derive an SE system for GLMs by applying the approximate message passing (AMP) algorithm.

Our contributions are summarized as follows:

- We perform statistical inference for GLMs in a high-dimensional setting. We also prove its asymptotic validity; specifically, the proposed confidence interval asymptotically attains the exact coverage probability. To this aim, we leverage the convex surrogate loss function and derive the SE system for a general class of GLMs from the AMP algorithm.

- We develop a moment-based estimator for hyperparameters in the SE system using an additive form of GLMs. We prove the consistency of the estimator for any $\kappa \in (0, \infty)$ under the restriction that the link function is asymmetric. Additionally, we verify that the estimator remains numerically stable even when κ is large.

1.1. Related Works and Comparison

Statistical estimation and inference in the high-dimensional regime (2) have been actively studied. For example, linear models have been considered by [BM11, EKBB$^+$13, TAH15, TAH18, TK18, EK18, MM21, HMRT22, MM22, BK25]; logistic regression models are by [SCC19, SC19], and others [TB23, Bel22, LGC$^+$21, SUI24]. In particular, [BKM$^+$19] deduced the Bayes-optimal estimation and generalization errors for GLMs. Additionally, [LGC$^+$21] provided a rigorous formula for the asymptotic training loss and generalization error in more general teacher-student settings. The relationship between our study and these two papers can be summarized as follows. First, our primary focus is on estimating unknown hyperparameters essential for statistical inference, rather than identifying the risk, which differentiates our motivation. Second, our proposed estimator with the convex surrogate loss is applicable to a wider class of MLEs, while its theoretical foundations have already been established in the prior research.

Regarding the estimation of SE hyperparameters, the first heuristic approach called *ProbeFrontier* was proposed by [SC19] though it is limited to logistic regression. This method estimates the signal strength (SS) appearing in the SE systems by utilizing the phase transition properties of the MLEs in logistic regression [CS20]. [YYMD21] developed the Signal Strength Leave-One-Out Estimator (SLOE) for the SS by adopting a different representation of the SE system for logistic regression. It is based on the non-regularized MLE, and therefore the estimation errors can diverge as κ increases and approaches to one. A subsequent paper [CLM24], appearing after the preprint release of our study, developed an estimator named into Method-of-Moment Inference (MoMI), by extending our moment-based inferenceby incorporating information from $\boldsymbol{X}$. This study also studied asymptotic variance of the coefficient estimator via the bootstrap method. A key difference from our results is that their method cannot handle the $\kappa \geq 1$ setting when a covariance matrix of $\boldsymbol{X}$ is unknown.

Table 1 summarizes the difference between our study and related methods. Accordingly, our method differs from existing approaches by achieving two key objectives: ensuring applicability even when κ is greater than 1 and enabling the estimation of hyperparameters beyond the SS needed for inference, e.g., variance of Gaussian noise. These are justified both theoretically and numerically. The intuitive reason our method can accommodate the $\kappa \geq 1$ setting, albeit at the cost of imposing an asymmetric assumption, is that it relies solely on the moment

Table 1
Comparison of methods based on κ values.

Method	Model	$\kappa < 1$	$\kappa \geq 1$	Condition on the link $g(\cdot)$
ProbeFrontier [SCC19]	Logistic	$\surd$		sigmoid $1/(1+\exp(-x))$
SLOE [YYMD21]	GLM	$\surd$		
MoMI [CLM24]	GLM	$\surd$		
Ours	GLM	$\surd$	$\surd$	restricted to asymmetric

information of Y. As a result, since our method does not use the information of $\boldsymbol{X}$, the inverse of the sample covariance matrix $(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top)^{-1}$ does not appear. This is a key advantage, as it ensures robustness against the instability of coefficient estimators and the inverse matrices as κ increases. On the other hand, a drawback of using only the moment information of Y is that SS cannot be identified when the link function is symmetric.

Without SE systems, [Bel22] introduced an inference framework that remains valid even when the link function is unknown under the identification condition $\mathrm{Var}(\boldsymbol{X}^\top\boldsymbol{\beta}) = 1$. However, when the link function is specified (i.e., the GLM case), this approach is not directly applicable since we cannot constrain SS to be 1. As for classification problems, [ML19, MLC19, SLCT21] established feasible methods for inference though they require prior knowledge of the covariance matrix of the covariates.

Various theoretical tools have been proposed for analyzing statistical models in the regime (2), including (i) AMP algorithms [DMM09, Bol14, BM11], (ii) convex Gaussian minimax theorem [TAH15, TAH18], (iii) leave-one-out techniques [EKBB+13], (iv) second-order Poincaré inequalities [Cha09], and (v) second-order Stein's formulae [BZ21, BS22]. In addition to (i), [Ran11] proposed generalized AMP applicable to GLMs. We use this to characterize the asymptotic behavior of an estimator of GLMs.

In a high-dimensional regime different from (2), statistical inferences for GLMs under sparse conditions have also been studied [GC16, JVDG16, SAH19, CGM21, LZCL23]. We note that the regime (2) admits the number of nonzero elements of coefficient s to be proportional to p, while classical high-dimensional statistics typically focus on sparse models with $s\log p = o(n)$.

1.2. State Evolution (SE)-based Confidence Interval: Logistic Regression Case

We review statistical inference frameworks that are valid in the high-dimensional regime (2), especially those aimed at confidence intervals. This framework for for a logistic regression is developed by [SC19, ZSC22]. Specifically, the studies showed that a properly corrected maximum likelihood estimator (MLE) is asymptotically normally distributed. Let $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^\top$ be the MLE. Then, as $n, p(n) \to \infty$ with $p(n)/n \to \kappa \in (0,1)$, we have the following for any

$j = 1, \dots, p$ satisfying $\sqrt{n}\tau_j\beta_j = O(1)$ if the MLE exists:

$$\frac{\sqrt{n}(\widehat{\beta}_j^{\mathrm{MLE}} - \mu\beta_j)}{\sigma/\tau_j} \xrightarrow{d} \mathcal{N}(0,1), \tag{3}$$

where $\tau_j^2 := \mathrm{Var}(X_{ij} \mid \boldsymbol{X}_{i\setminus j})$ is the conditional variance of the j-th element of $\boldsymbol{X}_i$ given $\boldsymbol{X}_{i\setminus j} := (X_{i,1}, \dots, X_{i,j-1}, X_{i,j+1}, \dots, X_{i,p})$. Here, (μ, σ) are the *state evolution* (SE) parameters that satisfy the following SE system,

$$\begin{cases} \kappa^2\sigma^2 & = \mathbb{E}_{(Q_1,Q_2)}\left[2g(Z)\left(\eta g(D)\right)^2\right], \\ 0 & = \mathbb{E}_{(Q_1,Q_2)}\left[g(Z)Z(\eta g(D))\right], \\ 1-\kappa & = \mathbb{E}_{(Q_1,Q_2)}\left[2g(Z)(1+\eta g'(D))^{-1}\right], \end{cases} \tag{4}$$

with $D = \mathrm{prox}_{\eta G}(\mu Z + \sqrt{\kappa}\sigma Q_2)$, $Z = \gamma Q_1$, and $(Q_1, Q_2) \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{I}_2)$. Here, the parameter $\gamma^2 = \lim_{n\to\infty} \mathrm{Var}(\boldsymbol{X}_i(n)^\top \boldsymbol{\beta}(n))$ is for *signal strength* (SS). The covariance structure of $\boldsymbol{X}_i$ is completely captured by γ^2 in this system. In this logistic regression case, the inverse link and the integrated inverse link functions are defined as $g(t) = 1/(1+\exp(-t))$ and $G(t) := \log(1+\exp(t))$, respectively. Here, $\mu, \sigma, \eta \in \mathbb{R}$ are the SE parameters. Remarkably, SE-based inference reduces the inference problem on high-dimensional coefficients to an SE system with a small number of scalar SE parameters.

The SE system is derived as a fixed-point equation of an iterative algorithm for an optimization problem on estimators. In the logistic regression case, we derive the SE system (4) for the MLE. The derivation is based on the approximate message passing, the details of which are provided in [FVRS22].

On the Basis of the asymptotic normality of the MLE in (3), we may construct a confidence interval for each β_j for $j = 1, \dots, p$, provided that the SE parameters are available. Specifically, given (μ, σ), the MLE-centric confidence interval with confidence level $1-\alpha \in (0,1)$ is formed as

$$\mathrm{CI}_{(1-\alpha)} := \left[\frac{\widehat{\beta}_j^{\mathrm{MLE}}}{\mu} - z_{(1-\alpha/2)}\frac{\sigma/\tau_j}{\sqrt{n}\mu}, \frac{\widehat{\beta}_j^{\mathrm{MLE}}}{\mu} + z_{(1-\alpha/2)}\frac{\sigma/\tau_j}{\sqrt{n}\mu}\right],$$

where $z_{\alpha/2}$ is the $(\alpha/2)$-quantile of a standard Gaussian distribution. Then we obtain $\Pr(\beta_j \in \mathrm{CI}_{(1-\alpha)}) \to 1-\alpha$ as $n, p(n) \to \infty$ in the sense of (2). Importantly, we need to know the SS parameter γ^2 in advance, because the SE parameters (μ, σ) depend on γ^2.

1.3. Notation

Define $\mathbb{R}_+ = (0, \infty)$. For a vector $\boldsymbol{b} = (b_1, \dots, b_p) \in \mathbb{R}^p$, b_j denotes a j-th element of $\boldsymbol{b}$ for $j = 1, \dots, p$. For function $f : \mathbb{R} \to \mathbb{R}$, $f'(\cdot)$ denotes the derivative of $f(\cdot)$. We say $f(\cdot)$ is *C-Lipschitz* with $C > 0$ if $f(\cdot)$ is Lipschitz continuous and its Lipschitz constant is C. For any convex function $f : \mathbb{R} \to \mathbb{R}$ and constant $\eta > 0$, we define the proximal operator $\mathrm{prox}_{\eta f} : \mathbb{R} \to \mathbb{R}$ as $\mathrm{prox}_{\eta f}(x) = \arg\min_{z\in\mathbb{R}} \left\{\eta f(z) + (x-z)^2/2\right\}$.

2. Preliminary

2.1. Generalized Linear Model

Suppose that we observe i.i.d. n pairs $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ of a feature vector $\boldsymbol{X}_i \equiv \boldsymbol{X}_i(n) \in \mathbb{R}^p$ and a target variable $Y_i \equiv Y_i(n) \in \mathcal{Y}$ that follow the GLM (1), where $\mathcal{Y}$ is a response space, such as $\mathbb{R}, \mathbb{R}_+, \{0,1\}, \{0,1,2,\dots\}$, and so on. Hereafter, we drop the dependence on n whenever it is clear from context. We assume that the feature vector is generated independently from $\boldsymbol{X}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ using the covariance matrix $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}(n) \in \mathbb{R}^{p \times p}$. The GLM can represent several models by specifying the inverse link function $g(\cdot)$ and distribution of Y_i for a given $\boldsymbol{X}_i$: for example, $Y_i \mid \boldsymbol{X}_i \sim \mathrm{Ber}(g(\boldsymbol{X}_i^\top \boldsymbol{\beta}))$ with $g(t) = 1/(1+\exp(-t))$ for the logistic regression model, and $Y_i \mid \boldsymbol{X}_i \sim \mathrm{Pois}(g(\boldsymbol{X}_i^\top \boldsymbol{\beta}))$ with $g(t) = \exp(t)$ for the Poisson regression model.

If a random variable $Y_i \mid \boldsymbol{X}_i$ has a density function (continuous case) or a mass function (discrete case) $f(\cdot \mid \boldsymbol{X}_i)$, then the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ is defined as $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}(n) = \arg\max_{b \in \mathbb{R}^p} \sum_{i=1}^n \log f(Y_i \mid \boldsymbol{X}_i)$. In a low-dimensional setting where $n \to \infty$ with fixed $p < \infty$, the MLE $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ asymptotically obeys the normal distribution as $\sqrt{n}(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}, \mathcal{I}_{\boldsymbol{\beta}}^{-1})$, where $\mathcal{I}_{\boldsymbol{\beta}}$ is the Fisher information matrix at the true coefficient $\boldsymbol{\beta}$. Once the matrix is estimated consistently, we can consider the inference of $\boldsymbol{\beta}$ conventionally. By contrast, this classical approach is no longer valid in a proportionally high-dimensional setting (2), as presented in [SC19].

2.2. Our Goal

Our goal is to construct a valid confidence interval for the unknown coefficient vector $\boldsymbol{\beta}$ of the GLM (1) that is valid in the high-dimensional regime (2). Specifically, for $\alpha \in (0,1)$ we construct a set $\mathrm{CI}_{(1-\alpha)} \subset \mathbb{R}$ such that we obtain $\Pr(\beta_j \in \mathrm{CI}_{(1-\alpha)}) \to 1-\alpha$ at the limit (2). Note that we need to introduce an estimator $\boldsymbol{\beta}$, since the MLE is not a valid instrument in the high-dimensional regime, as noted above.

3. Method

We present a confidence interval for GLMs in the high-dimensional regime by three steps, as overviewed in Section 1.2:

1. We construct a surrogate estimator for the coefficient $\boldsymbol{\beta}$ of GLM, and derive its SE systems.
2. We develop moment-based estimators for parameters in the derived SE system. This step constitutes the core of our novelty.
3. We construct a confidence interval based on the surrogate estimator and the moment-based estimator.

Each step is detailed in the following subsections.

3.1. Surrogate Estimator for Coefficient Vector $\boldsymbol{\beta}$

3.1.1. Definition

We develop an estimator for the true coefficient vector $\boldsymbol{\beta}$ by using a *surrogate loss* function [AHW95, AKK+14], which is characterized by $\ell : \mathbb{R}^p \times \mathbb{R}^p \times \mathcal{Y} \to \mathbb{R}$ with

$$\ell(\boldsymbol{b}; \boldsymbol{x}, y) := G\left(\boldsymbol{x}^\top \boldsymbol{b}\right) - y\boldsymbol{x}^\top \boldsymbol{b},$$

where $G(\cdot)$ is the function satisfying $G' = g$. Using the loss, we define the surrogate estimator

$$\widehat{\boldsymbol{\beta}}(n) := \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \ell(\boldsymbol{b}; \boldsymbol{X}_i, Y_i). \tag{5}$$

The use of a surrogate estimator is justified in two ways. First, if $g(\cdot)$ monotonically increases, the surrogate loss is convex in $\boldsymbol{b}$. This property provides computational and statistical advantages for a broader class of GLMs and plays a crucial role in the derivation of SE systems. Second, the minimizer of surrogate risk coincides with the true coefficient $\boldsymbol{\beta}$, which is expressed as follows:

Proposition 1 (Lemma 1 in [AKK+14])**.** *Consider an $\mathbb{R}^p \times \mathcal{Y}$-valued random element $(\boldsymbol{X}, Y)$ that follows GLM (1). If $g(\cdot)$ is increasing, then the coefficient vector $\boldsymbol{\beta}$ in (1) satisfies $\boldsymbol{\beta} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \mathbb{E}\left[\ell(\boldsymbol{b}; \boldsymbol{X}, Y) \mid \boldsymbol{X}\right]$.*

This claims that the surrogate loss estimator is a reasonable extension of MLE for GLMs. Modeling the distribution of $Y_i \mid \boldsymbol{X}_i$ to follow an exponential dispersion family [Jør87] and choosing $g(\cdot)$ as the inverse of the canonical link function make the surrogate estimator (5) equivalent to the MLE, covering logistic and Poisson regression.

3.1.2. SE System for Surrogate Estimator of GLMs

We derive an SE system associated with the surrogate estimator (5). As this estimator is given by the minimization problem of the convex loss function, we can derive the SE system for the estimator, as described in Section 1.2. Recall that $G(\cdot)$ is a function satisfying $G' = g$, and $\gamma^2 = \lim_{n\to\infty} \mathrm{Var}(\boldsymbol{X}_i(n)^\top \boldsymbol{\beta}(n))$ is the signal strength (SS) parameter.

Proposition 2. *Suppose that the model (1) can be rewritten as $Y = h(\boldsymbol{X}^\top \boldsymbol{\beta}, \varepsilon)$ with $\mathbb{R}$-valued non-random function $h(\cdot,\cdot)$ and the noise variable $\varepsilon \in \mathbb{R}$ independent of $\boldsymbol{X}$. If $\kappa \in (0, 1)$ and the surrogate estimator in (5) is bounded, then the corresponding SE system is given by*

$$\begin{cases} \kappa^2\sigma^2 & = \eta^2 \mathbb{E}_{(Q_1,Q_2,\bar{Y})}\left[\left(\bar{Y} - g\left(D\right)\right)^2\right], \\ 0 & = \mathbb{E}_{(Q_1,Q_2,\bar{Y})}\left[Z\left(\bar{Y} - g\left(D\right)\right)\right], \\ 1-\kappa & = \mathbb{E}_{(Q_1,Q_2,\bar{Y})}\left[\left(1 + \eta g'\left(D\right)\right)^{-1}\right], \end{cases} \tag{6}$$

where $D = \mathrm{prox}_{\eta G}(\mu Z + \sqrt{\kappa}\sigma Q_2 + \eta \bar{Y})$, $Z = \gamma Q_1$, $(Q_1, Q_2) \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{I}_2)$, *and* $\bar{Y} = h(Z, \varepsilon)$ *with a random variable* $\varepsilon \in \mathbb{R}$ *independent of* Z.

In view of (6), the SE parameters (μ, σ^2, η) depend on three components: (i) $\kappa = \lim_{n\to\infty} p(n)/n$, (ii) the SS parameter $\gamma^2 = \lim_{n\to\infty} \mathrm{Var}(\boldsymbol{X}_i(n)^\top \boldsymbol{\beta}(n))$, and (iii) the distribution of $\bar{Y}$ determined by the model. Some numerical plots of the equations (6) will be presented in Section A.1.

Deriving the parameters of this equation is a non-trivial task. Although we can set $\kappa = p(n)/n$ and regard it as known, the estimation of γ^2 and other parameters (if present) of the distribution of $\bar{Y}$ for each GLM is non-trivial. In the next subsection, we propose a method for estimating them.

Remark 1. *If the distribution of* $Y \mid \boldsymbol{X}$ *has a single parameter (e.g., Bernoulli, Poisson, and exponential distribution), then one can specify* $\varepsilon \sim \mathrm{Unif}[0,1]$ *and the model assumption* $\mathbb{E}[Y \mid \boldsymbol{X}] = g(\boldsymbol{X}^\top \boldsymbol{\beta})$ *implies that* γ^2 *fully characterizes the SE parameters* (μ, σ^2, η). *Thus, in this case, the SS parameter* γ^2 *is the only unknown parameter to be estimated. Meanwhile, estimating the distribution of* $Y \mid \boldsymbol{X}$ *is difficult if it has multiple parameters. Even in such cases, we propose a method for their estimation, focusing on the case of* $Y \mid \boldsymbol{X}$ *following a normal distribution; see Section 3.2.2.*

3.2. Moment-Based Estimation for SE Parameter

We propose a novel estimator for the SS parameter γ^2 and other parameters of the distribution of $\bar{Y}$ in (6) to access the SE parameters (μ, σ^2) for inference. Once we estimate γ^2 and the parameters of $\bar{Y}$, we can solve system (6) by substituting the estimators to obtain the SE parameters.

3.2.1. Additive Form of GLM

We convert the GLM (1) into the *additive model* with observations $(Y_i, \boldsymbol{X}_i)$:

$$Y_i = g(\boldsymbol{X}_i^\top \boldsymbol{\beta}) + e_i \quad \text{with} \quad e_i = Y_i - \mathbb{E}[Y_i \mid \boldsymbol{X}_i], \quad i = 1, \ldots, n, \tag{7}$$

where e_i is a noise variable that satisfies $\mathbb{E}[e_i \mid \boldsymbol{X}_i] = 0$. Note that the parameters of the distribution of $e_i \mid \boldsymbol{X}_i$ may depend on $\boldsymbol{X}_i$, $i = 1, \ldots, n$. While the original GLM (1) highlights the conditional mean, the additive model (7) provides a more analyzable form of the distribution by introducing the variable e_i. This structure is essential for our following estimation method.

3.2.2. Moment-Based Estimation when $Y \mid \mathbf{X}$ is Gaussian

We first apply our moment-based method to the Gaussian output: $Y \mid \boldsymbol{X} \sim \mathcal{N}(g(\boldsymbol{X}^\top \boldsymbol{\beta}), \sigma_e^2)$ with an unknown variance parameter $\sigma_e^2 > 0$. This is a typical case of GLMs including a nonlinear regression setup, but it has additional unknown parameter σ_e^2 to be estimated, as well as the SS parameter γ. We first consider this rather complicated case where we estimate both γ and σ_e^2 as it serves as an appropriate introduction to our approach.

At the beginning, we characterize the parameters γ^2 and σ_e^2 by the moments of Y. In the limit, the additive model (7) has the form,

$$\bar{Y} = g(Z_\gamma) + \bar{e} \quad \text{with} \quad Z_\gamma \sim \mathcal{N}(0,\gamma^2) \quad \text{and} \quad \bar{e} \sim \mathcal{N}(0,\sigma_e^2), \tag{8}$$

where Z_γ and $\bar{e}$ are independent. Then, we obtain the condition with second and fourth moments as

$$\Psi(\gamma,\sigma_e) := \begin{bmatrix} \mathbb{E}[\bar{Y}^2] - \mathbb{E}[g(Z_\gamma)^2] - \sigma_e^2 \\ \mathbb{E}[\bar{Y}^4] - \mathbb{E}[g(Z_\gamma)^4] - 6\mathbb{E}[g(Z_\gamma)^2]\sigma_e^2 - 3\sigma_e^4 \end{bmatrix} = \mathbf{0} \in \mathbb{R}^2. \tag{9}$$

This simultaneous equation characterizes the parameter (γ,σ_e^2) as a solution. The uniqueness of the solution to this system is validated numerically. The solution is identifiable for any $g(\cdot)$, except when $g(\cdot)$ is a constant function.

We consider estimation of (γ,σ_e^2) using an empirical analog of the equation (9). Since $g(\cdot)$ is known, we can simulate $h_2(\varsigma) := \mathbb{E}[g(Z_\varsigma)^2]$ and $h_4(\varsigma) := \mathbb{E}[g(Z_\varsigma)^4]$ by generating $Z_\varsigma \sim \mathcal{N}(0,\varsigma^2)$ for each $\varsigma > 0$. Using the observations $\{Y_i\}_{i=1}^n$, we then define the estimators as the solution to

$$(\widehat{\gamma},\widehat{\sigma}_e) := \left\{(\varsigma,\varsigma_e) \in \mathbb{R}_+^2 : \Psi_n(\varsigma,\varsigma_e) = \mathbf{0}\right\}, \tag{10}$$
$$\text{where } \Psi_n(\varsigma,\varsigma_e) = \begin{bmatrix} n^{-1}\sum_{i=1}^n Y_i^2 - h_2(\varsigma) - \varsigma_e^2 \\ n^{-1}\sum_{i=1}^n Y_i^4 - h_4(\varsigma) - 6h_2(\varsigma)\varsigma_e^2 - 3\varsigma_e^4 \end{bmatrix} \in \mathbb{R}^2.$$

The equation in (10) can be viewed as an empirical analogue of (9), provided that the convergence $n^{-1}\sum_{i=1}^n Y_i^a \xrightarrow{\text{a.s.}} \mathbb{E}[\bar{Y}^a]$ for $a \in \{2,4\}$ is true under some assumptions. The solution is obtained by root-finding algorithms, such as the Gauss-Newton method.

An advantage of the moment-based estimator is its independence from both the estimator $\widehat{\boldsymbol{\beta}}$ and an inverse of high-dimensional matrix $(\mathbb{X}^\top\mathbb{X})^{-1}$ with $\mathbb{X} = (\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)^\top$, that diverge as κ increases. Hence, the moment-based estimator is stable with any $\kappa \in (0,\infty)$. Additionally, since this method relies only on observations through Y, it is computationally efficient.

3.2.3. Moment-Based Estimation for Other Cases

We next study moment-based estimation in other situations of GLMs with a single parameter. The problem becomes simpler since we only need to estimate γ, unlike the Gaussian case. Examples are given in Remark 1.

In this case, we only consider the first moment using the limit form (8):

$$\tilde{\Psi}(\gamma) := \mathbb{E}[\bar{Y}] - \mathbb{E}[g(Z_\gamma)] = 0, \tag{11}$$

where we have used the fact that $\mathbb{E}[e_1] = \mathbb{E}[\mathbb{E}[e_1 \mid \boldsymbol{X}_1]] = 0$ by the assumption. The estimator of γ is obtained as a solution to the linear equation with $h_1(\varsigma) := \mathbb{E}[g(Z_\varsigma)]$:

$$\widehat{\gamma} := \left\{\varsigma \in \mathbb{R}_+ : \tilde{\Psi}_n(\varsigma) = 0\right\}, \quad \text{where } \tilde{\Psi}_n(\varsigma) := n^{-1}\sum_{i=1}^n Y_i - h_1(\varsigma). \tag{12}$$

Similarly to Section 3.2.2, this estimator is stable for any $\kappa \in (0, \infty)$ and computationally efficient. If necessary, we can construct an estimator considering higher-order moments as in Section 3.2.2.

3.3. Confidence Interval

We construct a confidence interval for $\boldsymbol{\beta}$ by using the estimators in Sections 3.2 and $\widehat{\boldsymbol{\beta}}$ in (5). Let $(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\eta})$ be the solutions to the SE system (6), with γ^2 replaced by $\widehat{\gamma}^2$. We also introduce an estimator $\widehat{\tau}_j$ of the conditional variance τ_j^2, the details of which are provided in Section A.2. Then, our confidence interval with a confidence level $1 - \alpha \in (0, 1)$ is defined as:

$$\mathcal{CI}_{1-\alpha,j} := \left[\frac{\widehat{\beta}_j}{\widehat{\mu}} - z_{(1-\alpha/2)} \frac{\widehat{\sigma}}{\sqrt{n}\widehat{\mu}\widehat{\tau}_j}, \frac{\widehat{\beta}_j}{\widehat{\mu}} + z_{(1-\alpha/2)} \frac{\widehat{\sigma}}{\sqrt{n}\widehat{\mu}\widehat{\tau}_j} \right], \ j = 1, \ldots, p. \tag{13}$$

4. Theory

We show the theoretical validity of each of the estimators obtained above. Specifically, we prove the asymptotic normality of the surrogate estimator and the consistency of the moment-baesd estimator, showing the validity of the confidence interval.

4.1. Asymptotic Normality of Surrogate Estimator

We derive the theoretical results of the surrogate estimator $\widehat{\boldsymbol{\beta}}$. First, we give the following assumptions.

Assumption 1. *We consider the following conditions:*

(A1) $\boldsymbol{X}$ *is generated as* $\boldsymbol{X} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$*, and* $\mathrm{Var}(\boldsymbol{X}^\top \boldsymbol{\beta})$ *has a convergence limit* $\gamma^2 < \infty$ *as* $n \to \infty$*.*

(A2) An inverse link function $g : \mathbb{R} \to \mathbb{R}$ *is monotonically increasing and L-smooth (i.e., the derivative of* $g(\cdot)$ *is L-Lipschitz continuous).*

(A3) We can write a GLM as $Y = h(\boldsymbol{X}^\top \boldsymbol{\beta}, \varepsilon)$ *with a non-random function* $h : \mathbb{R}^2 \to \mathbb{R}$ *and an* $\mathbb{R}$*-valued random variable* ε *independent of* $\boldsymbol{X}$*, such that* ε *has a finite second moment and* $h(\cdot, \cdot)$ *is Lipschitz continuous with respect to the first argument.*

(A4) There exists a positive solution to the SE system (6).

Convergence of $\mathrm{Var}(\boldsymbol{X}^\top \boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}$ in (A1) is a constraint on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. (A3) is a technical requirement for the Lipschitzness of SE parameters with respect to γ^2. Examples of $h(\cdot, \cdot)$ are provided in Section A.6. In the sequel, we focus only on the situation in which the estimator (5) exists asymptotically almost surely; that is, $\lim_{n \to \infty} \|\widehat{\boldsymbol{\beta}}\| < \infty$ a.s. If this does not hold, we may consider regularization to guarantee its existence; see Section A.5 for this point.

We obtain the asymptotic normality of the adjusted test statistics for each coordinate of $\widehat{\boldsymbol{\beta}}$ with the oracle SE parameters μ and σ^2.

Proposition 3. *Suppose that we know a solution to the system of nonlinear equations* (6)*, and the estimator* (5) *almost surely exists asymptotically. Under (A1), (A2), and (A4) in Assumption 1, as $n, p(n) \to \infty$ and $p(n)/n \to \kappa \in (0,1)$, we have the following for $j = 1, \dots, p(n)$:*

$$\frac{\sqrt{n}(\widehat{\beta}_j - \mu\beta_j)}{\sigma/\tau_j} \xrightarrow{d} \mathcal{N}(0,1).$$

This is the first result to demonstrate the marginal asymptotic normality of estimators for GLMs in a high-dimensional setting, including MLEs with canonical links. This differs from the classical result $\sqrt{n}(\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}_{\boldsymbol{\beta}}^{-1})$ in a setup with $p < \infty$. We regard the marginal convergence as an extension of [ZSC22] on MLEs for logistic regression to more general models. A feasible construction of the asymptotic normality is provided in Proposition 4.

4.2. Consistency of Moment-Based Estimator

We show the consistency of the estimators for the SS parameters in the cases of Sections 3.2.2 and 3.2.3. We need the following assumption.

Assumption 2. *We consider the following conditions:*

(A5) For any $\epsilon > 0$, we have both $\inf_{(\varsigma,\varsigma_e):|\varsigma-\gamma|+|\varsigma_e-\sigma_e|>\epsilon} \|\Psi(\varsigma,\varsigma_e)\| > 0$ and $\inf_{\varsigma:|\varsigma-\gamma|>\epsilon} |\tilde{\Psi}(\varsigma)| > 0$.

(A6) There exist constants $C, c, \epsilon > 0$ such that $|g(z)| \leq \exp(Cz^{2-\epsilon} + c)$ for any $z \in \mathbb{R}$.

Condition (A5) identifies the SS parameters (γ, σ_e^2) by the moment equations. This condition is not satisfied if the moments of Y do not have sufficient information, such as the case of logistic regression. However, a later study [CLM24] proposes a method that utilizes information from $\boldsymbol{X}$ to handle logistic regression using our moment-based technique especially when $n > p$. We also discuss a sufficient condition for (A5) in Section A.8. The condition (A6) is required for the existence of $\mathbb{E}[g(Z_\gamma)^q]$ for $q \geq 1$ used in the equations (8) and (11).

We then obtain the following results for the consistency of the estimators for the SS parameters:

Theorem 1. *Under Assumption 1 (A1) and Assumption 2, $\widehat{\gamma}$ defined in* (12) *satisfies $\widehat{\gamma}^2 \xrightarrow{\text{a.s.}} \gamma^2$ and $(\widehat{\gamma}^2, \widehat{\sigma}_e^2)$ defined in* (10) *satisfies $(\widehat{\gamma}^2, \widehat{\sigma}_e^2) \xrightarrow{\text{a.s.}} (\gamma^2, \sigma_e^2)$, as $n, p(n) \to \infty$ and $p(n)/n \to \kappa \in (0, \infty)$.*

4.3. Asymptotic Validity of Proposed Confidence Interval

At last, we demonstrate the asymptotic validity of our bias correction for statistical inference. To observe this, we use the asymptotic normality with the estimated SE parameters.

Proposition 4. *Suppose that all the conditions in Assumptions 1 and 2 hold, the estimator (5) almost surely exists asymptotically, and $\widehat{\tau}_j^2$ is a consistent estimator of the conditional variance τ_j^2. Then, for any confidence level $(1-\alpha) \in (0,1)$ and for every $j = 1, \dots, p(n)$ satisfying $\sqrt{n}\tau_j\beta_j = O(1)$, we obtain the following as $n, p(n) \to \infty$, where $p(n)/n \to \kappa \in (0,1)$:*

$$\frac{\sqrt{n}(\widehat{\beta}_j - \widehat{\mu}\beta_j)}{\widehat{\sigma}/\widehat{\tau}_j} \xrightarrow{d} \mathcal{N}(0,1).$$

To the best of our knowledge, this is the first result establishing the asymptotic normality with the estimated SE parameters in the GLM literature. Its proof relies on the Lipschitz continuity of SE estimators with respect to SS. Consequently, the proposed confidence interval asymptotically achieved a confidence level $(1-\alpha)$.

Theorem 2. *Assume the settings of Proposition 4. Then, for any confidence level $(1-\alpha) \in (0,1)$ and for every $j = 1, \dots, p(n)$ satisfying $\sqrt{n}\tau_j\beta_j = O(1)$, we obtain $\Pr(\beta_j \in \mathcal{CI}_{1-\alpha,j}) \to 1-\alpha$, as $n, p(n) \to \infty$ where $p(n)/n \to \kappa \in (0,1)$.*

5. Experiment

5.1. Empirical Performance of $\widehat{\gamma}$ and $\widehat{\sigma}_e^2$

We fix $n = 4000$ and vary $p = \kappa n$ over $\kappa \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. The SS parameter is also fixed at $\gamma^2 = 1$. We independently generate n realizations of the feature vector $\boldsymbol{X} \in \mathbb{R}^p$ from the p-variate normal distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}_{ij} = 0.5^{|i-j|}$. The true regression coefficients β_j for $j = 1, \dots, p$ are independently drawn from a normal distribution with the variance determined by γ^2. We calculate $g(\boldsymbol{X}^\top\boldsymbol{\beta})$ conditional on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ and finally draw the response Y from a given distribution.

We consider three cases to check the performance of $\widehat{\gamma}$ in (12): (i) Poisson regression $g(t) = \exp(t)$, $Y \mid \boldsymbol{X} \sim \mathrm{Pois}(g(\boldsymbol{X}^\top\boldsymbol{\beta}))$, (ii) piecewise regression $g(t) = \min(5t, 0.1t)$, $Y \mid \boldsymbol{X} \sim \mathcal{N}(g(\boldsymbol{X}^\top\boldsymbol{\beta}), 0.04)$, and (iii) complementary log-log (cloglog) regression $g(t) = 1 - \exp(-\exp(t))$, $Y \mid \boldsymbol{X} \sim \mathrm{Bern}(g(\boldsymbol{X}^\top\boldsymbol{\beta}))$. We further investigate two situations to evaluate the behavior of $\widehat{\sigma}_e^2$ in (10): (i) piecewise regression $g(t) = \min(5t, 0.1t)$, $Y \mid \boldsymbol{X} \sim \mathcal{N}(g(\boldsymbol{X}^\top\boldsymbol{\beta}), 0.04)$ and (ii) squared regression $g(t) = t^2$, $Y \mid \boldsymbol{X} \sim \mathcal{N}(g(\boldsymbol{X}^\top\boldsymbol{\beta}), 0.04)$. These results are summarized in Figure 1; the upper and lower panels demonstrate stability and consistency of $\widehat{\gamma}$ and $\widehat{\sigma}_e^2$, respectively, regardless of the values of κ.

5.1.1. Comparison with Other Methods.

We compare the moment-based SS estimator $\widehat{\gamma}^2$ with the previous leave-one-out-based SS estimator, SLOE $\tilde{\gamma}^2_{\mathrm{SLOE}}$ [YYMD21]. Since the original SLOE is designed for logistic regression, we extend it to GLMs; see Section A.4 for details. Here, note that the estimand of SLOE is not exactly SS but rather the

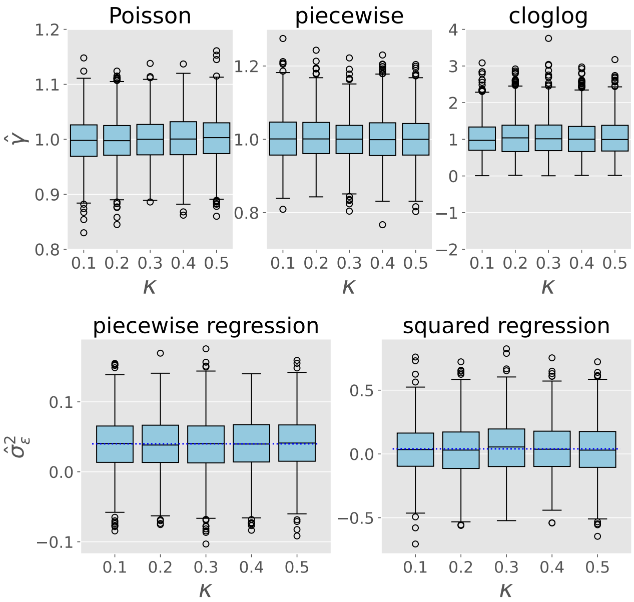

Fig 1: Numerical performance of $\widehat{\gamma}$ and $\widehat{\sigma}_e^2$ over 1000 simulations. The true values are $\gamma = 1$ and $\sigma_\varepsilon^2 = 0.04$.

corrupted signal strength, $\gamma_c^2 = \widehat{\boldsymbol{\beta}}^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{\beta}}$. In the setting with $n = 1000$ and $\gamma = 2$, we plot the scaled squared estimation errors $(\widehat{\gamma}/\gamma - 1)^2$ and $(\tilde{\gamma}_{\mathrm{SLOE}}/\gamma_c - 1)^2$, respectively, over 100 replications. We consider the Poisson regression and piecewise regression as in Section 5.1.

Figure 2 indicates that our moment-based estimator performs well even when κ exceeds 1 in both cases while SLOE diverges as κ increases and approaches 1. Additionally, regarding the computational load, SLOE requires a longer computation time as κ increases; whereas our proposed method remains nearly invariant.

5.2. *Coverage Proportion of Proposed Confidence Interval*

We compare the empirical coverage proportion of the proposed confidence interval (13) with that of the classical CI with $n = 1000$, $\kappa \in \{0.1, 0.2, \dots, 0.5\}$, and $\gamma = 1$. The proposed CI is calculated by solving a system of nonlinear equations using the estimated SS. The classical CI for β_j, $j = 1, \dots, p$, is constructed as $[\widehat{\beta}_j \pm z_{(1-\alpha/2)}(\widehat{\mathcal{I}}_{jj}^{-1}(\widehat{\boldsymbol{\beta}})/n)^{1/2}]$, where $\widehat{\mathcal{I}}(\widehat{\boldsymbol{\beta}}) = n^{-1}\sum_{i=1}^n g'(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})\boldsymbol{X}_i\boldsymbol{X}_i^\top$ is the empirical Fisher information matrix evaluated at $\widehat{\boldsymbol{\beta}}$. We analyze the Poisson regression model. Figure 3 illustrates that our proposed interval achieves theoretical coverage in all cases; whereas the coverage of the conventional method decreases as κ increases.

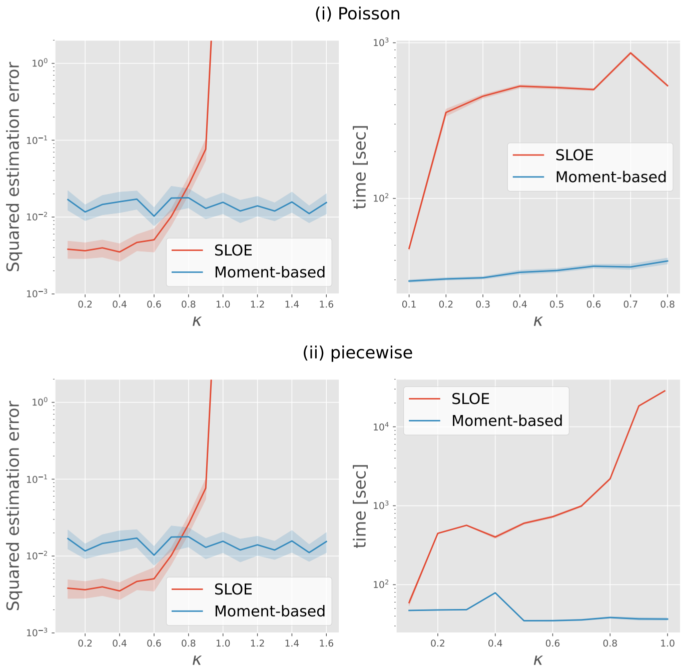

Fig 2: Scaled squared estimation error of SLOE and our proposed method, in the piecewise regression. The LOO-based method is our extension of the existing method; see Section A.4.

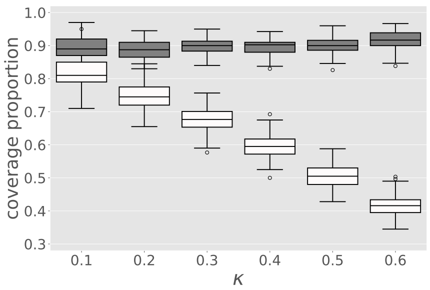

Fig 3: Coverage proportion of true coefficient with 90% classical (white) and proposed (gray) CIs over 100 simulations with $n = 1000$.

5.3. Real Data Application

We consider an application for the Cleveland Clinic Heart Disease dataset [DJS+89] from the UCI Machine Learning Repository [DG17]. This contains 303 observations of 14 variables. The target is the presence of heart disease, which is integer-valued from 0 (no presence) to 4. To realize a setting with a high-dimensional regime, for each $i = 1, \dots, 303$, we generate an independent random vector from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_{198})$ and concatenate it with the original 14 variable to obtain the covariate vector $\boldsymbol{X}_i$. Hence, we have $\kappa = (14 + 198)/303 \approx 0.7$. We constructed CIs with $\alpha = 0.1$ for $\boldsymbol{X}_i^\top \boldsymbol{\beta}$ by our proposed CI and a classical CI by the MLE for Poisson regression; see Section A.7 for the construction of the classical CI. We approximate the coverage proportion of CIs to $\boldsymbol{X}_i^\top \boldsymbol{\beta}$ using Y_i, and preclude samples with responses that take zero because they cannot be covered.

As a result, the coverage proportion of classical CIs is $139/139 = 100\%$ and that of the proposed CIs is $126/139 \approx 90.6\%$. Given that the ideal coverage proportion is $1 - \alpha = 0.9$, the proposed CI provides more appropriate coverage though the classical CI is too conservative. This result implies that classical CIs overfit the samples and underestimate the uncertainty of the estimation, while the proposed CIs control the coverage rate nearly at the preassigned level and can evaluate the uncertainty much more accurately.

6. Conclusion

We have developed a statistical inference method for GLMs in high dimensions. Our approach extends the SE-based inference method designed for a logistic regression model to GLMs. Specifically, we have proposed a surrogate estimator for GLMs, an associated SE system, and a method to estimate the necessary parameters in the system. Our methodology works well in terms of both theory and experimentation. One limitation of our method is that it requires Gaussianity of the covariate; however, this limitation can be relaxed by applying studies on the universality of methodologies in terms of data, for example, [MS22, VKM22].

Appendix A: Supportive Information

A.1. Numericals of SE System

We demonstrate the solutions to the system for Poisson regression in Figure 4 and piecewise regression in Figure 5. These plots are numerical solutions to the SE system over 100 simulations with their mean and 95% bootstrap confidence intervals.

A.2. Construction of $\widehat{\tau}_j^2$

To construct the corrected confidence interval (13) with correlated features, we have to estimate the conditional variance parameter, $\tau_j^2 = \mathrm{Var}(\boldsymbol{X}_{ij} | \boldsymbol{X}_{i \setminus j})$, for

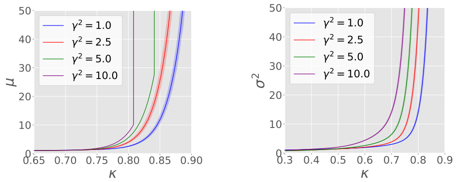

Fig 4: Computed bias μ (left) and variance σ^2 (right) for Poisson regression model $Y \mid \boldsymbol{X} \sim \mathrm{Pois}(g(\boldsymbol{X}^\top \boldsymbol{\beta}))$ with $g(t) = \exp(t)$.

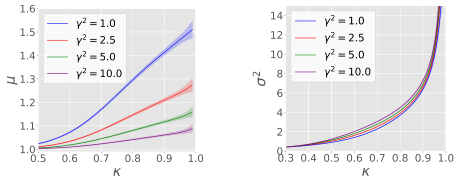

Fig 5: Computed bias μ (left) and variance σ^2 (right) for regression model $Y \mid \boldsymbol{X} \sim \mathcal{N}(g(\boldsymbol{X}^\top \boldsymbol{\beta}), 0.2)$ with $g(t) = \min(5t, 0.1t)$.

each $j = 1, \dots, p$. In this article, we follow [ZSC22]. They consider the residual sum of squares RSS_j obtained by regression of $\boldsymbol{X}_j = (\boldsymbol{X}_{1j}, \dots, \boldsymbol{X}_{nj})^\top \in \mathbb{R}^n$ onto a sub-vector of the input $\boldsymbol{X}_{\setminus j} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, \dots, \boldsymbol{X}_p) \in \mathbb{R}^{n \times (p-1)}$. Owing to the Gaussianity of $\boldsymbol{X}_i$, it satisfies

$$\mathrm{RSS}_j = \boldsymbol{X}_j^\top \boldsymbol{P}_{\boldsymbol{X}_{\setminus j}}^\perp \boldsymbol{X}_j \sim \tau_j^2 \chi^2_{n-p+1},$$

where $\boldsymbol{P}_{\boldsymbol{X}_{\setminus j}}^\perp$ is the orthogonal projection matrix onto the orthogonal complement of column space spanned by $\boldsymbol{X}_{\setminus j}$. Then, we immediately obtain an unbiased estimator of τ_j^2:

$$\widehat{\tau}_j^2 = \frac{1}{n-p+1} \mathrm{RSS}_j.$$

A.3. Master Theorem of Generalized Approximate Message Passing (GAMP)

We present the GAMP algorithm by [Ran11] and its associated theoretical result, which are key tools to inspect the limiting distributional behavior of the estimator $\widehat{\boldsymbol{\beta}}$.

First, we provide the GAMP algorithm, which generate a sequence of parameters $\tilde{\boldsymbol{\beta}}^k$ for an index $k \in \mathbb{N} \cup \{0\}$ based on the minimization problem in (5) and a limit of the sequence corresponds to the estimator $\widehat{\boldsymbol{\beta}}$. Let $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)^\top \in \mathbb{R}^{n \times p}$, $\breve{\boldsymbol{X}} = \boldsymbol{X}/\sqrt{n}$, and $\boldsymbol{Y} = (Y_1, \dots, Y_n) \in \mathbb{R}^n$. Given initial values $\bar{\eta}_0 > 0, \tilde{\boldsymbol{\beta}}^0$, and $\tilde{\boldsymbol{\xi}}^0 = \breve{\boldsymbol{X}} \tilde{\boldsymbol{\beta}}^0$, a GAMP recursion takes the form, for each $k \in \mathbb{N} \cup \{0\}$,

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}^{k+1} &= \frac{\bar{\eta}_{k+1}}{\kappa} \breve{\boldsymbol{X}}^\top \left\{ \boldsymbol{Y} - g\left(\mathrm{prox}_{\bar{\eta}_k G}(\tilde{\boldsymbol{\xi}}^k + \bar{\eta}_k \boldsymbol{Y}) \right) \right\} + \frac{\bar{\eta}_{k+1}}{\bar{\eta}_k} \tilde{\boldsymbol{\beta}}^k, \\
\tilde{\boldsymbol{\xi}}^{k+1} &= \breve{\boldsymbol{X}} \tilde{\boldsymbol{\beta}}^{k+1} - \bar{\eta}_{k+1} \left\{ \boldsymbol{Y} - g\left(\mathrm{prox}_{\bar{\eta}_k G}(\tilde{\boldsymbol{\xi}}^k + \bar{\eta}_k \boldsymbol{Y}) \right) \right\}.
\end{aligned}$$

Here, $\bar{\eta}_k$ is updated with $\bar{\mu}_k$ and $\bar{\sigma}_k^2$ as following:

$$\begin{aligned}
\bar{\eta}_{k+1} &= \kappa\bar{\eta}_k \left(1 - \mathbb{E}_{(Q_1,Q_2,U)}\left[\frac{1}{1+\bar{\eta}_k g'\left(d_k\right)}\right]\right)^{-1}, \\
\bar{\mu}_{k+1} &= \frac{\bar{\eta}_{k+1}}{\gamma^2}\mathbb{E}_{(Q_1,Q_2,U)}\left[\gamma Q_1 \left\{h(\gamma Q_1, U) - g\left(d_k\right)\right\}\right], \\
\bar{\sigma}_{k+1}^2 &= \frac{\bar{\eta}_{k+1}^2}{\kappa^2}\mathbb{E}_{(Q_1,Q_2,U)}\left[\left\{h(\gamma Q_1, U) - g\left(d_k\right)\right\}^2\right],
\end{aligned}$$

where

$$\begin{aligned}
&d_k = \operatorname{prox}_{\bar{\eta}_k(\gamma)G}(\bar{\mu}_k\gamma Q_1 + \sqrt{\kappa}\bar{\sigma}_k Q_2 + \eta_k h(\gamma Q_1, U)), \\
&(Q_1, Q_2) \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{I}_2), \qquad U \perp\!\!\!\perp (Q_1, Q_2).
\end{aligned}$$

Here, $U \in \mathbb{R}$ is a random variable independent of (Q_1, Q_2). This algorithm is proposed by [Ran11]. It is closely related to the linearized alternating direction method of multiplier [RSR$^+$16].

Second, we provide the following result for the estimator $\widehat{\boldsymbol{\beta}}$, which is regarded as a limit of the parameter sequence generated by the GAMP algorithm. The result follows the master theorem for the GAMP algorithm presented by [FVRS22].

Lemma 1. *Assume that $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ is an i.i.d. sample, and $\boldsymbol{X}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{I}_p)$ independent of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$. For $r \in [2, \infty)$, suppose that the empirical distributions $p^{-1}\sum_{j=1}^p \delta_{\sqrt{n}\beta_j}$ and $n^{-1}\sum_{i=1}^n \delta_{\varepsilon_i}$ converge in the r-Wasserstein sense to the distributions of $\bar{\beta}$ and $\bar{\varepsilon}$ with finite r-th order moment, respectively. Let $Z \sim \mathcal{N}(0, \gamma^2)$ be independent of $\bar{\varepsilon}$, and $G, \tilde{G} \sim \mathcal{N}(0, 1)$ be independent of $\bar{\beta}$ and Z. Then, for finite $\widehat{\boldsymbol{\beta}}$ defined in (5) and any pseudo-Lipschitz function $\psi : \mathbb{R}^2 \to \mathbb{R}$ and $\tilde{\psi} : \mathbb{R}^3 \to \mathbb{R}$ of order r[1], we have*

$$\begin{aligned}
\frac{1}{p}\sum_{j=1}^p \psi\left(\sqrt{n}\widehat{\beta}_j, \sqrt{n}\beta_j\right) &\xrightarrow{\text{a.s.}} \mathbb{E}\left[\psi\left(\mu\bar{\beta} + \sigma G, \bar{\beta}\right)\right], \\
\frac{1}{n}\sum_{i=1}^n \tilde{\psi}\left(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}, \boldsymbol{X}_i^\top\boldsymbol{\beta}, \varepsilon_i\right) &\xrightarrow{\text{a.s.}} \mathbb{E}\left[\tilde{\psi}\left(\operatorname{prox}_{\eta\ell}\left(\mu_Z Z + \sigma_Z\tilde{G}\right), Z, \bar{\varepsilon}\right)\right],
\end{aligned}$$

as $n, p(n) \to \infty$ with $p(n)/n \to \kappa \in (0, 1)$.

The convergence of pseudo-Lipschitz functions of order 1 is equivalent to the convergence of the 1-Wasserstein distance between the empirical distributions of each coordinate, as established through Kantorovich-Rubinstein duality [KR58].

Proof of Lemma 1. We prove this lemma by the general proof strategy discussed in Section 4.4 in [FVRS22]. It consists of three steps: (i) find a fixed point of the GAMP recursion; (ii) consider the stationary version of the GAMP recursion;

[1] A function $\psi : \mathbb{R}^m \to \mathbb{R}$ is said to be pseudo-Lipschitz of order r if there exists a constant $L > 0$ such that for any $t_0, t_1 \in \mathbb{R}^m$, $\|\psi(t_0) - \psi(t_1)\|_2 \le L(1 + \|t_0\|_2^{r-1} + \|t_1\|_2^{r-1})\,\|t_0 - t_1\|$.

(iii) show that the stationary version of the GAMP iterate converges to the estimator $\widehat{\boldsymbol{\beta}}$. Following the strategy, the step (i) and the step (ii) are simply achieved by the convexity and smoothness of the surrogate loss function by the assumption. Given the fixed point $\bar{\eta}_*$ of $\bar{\eta}_k$, and initial values $\boldsymbol{\beta}^0, \boldsymbol{\xi}^0 = \breve{\boldsymbol{X}}\boldsymbol{\beta}^0$, the stationary version of the GAMP algorithm takes the form, for each $k \in \mathbb{N} \cup \{0\}$,

$$\boldsymbol{\beta}^{k+1} = \frac{\bar{\eta}_*}{\kappa}\breve{\boldsymbol{X}}^\top \left\{\boldsymbol{Y} - g\left(\operatorname{prox}_{\bar{\eta}_* G}(\boldsymbol{\xi}^k + \bar{\eta}_*\boldsymbol{Y})\right)\right\} + \boldsymbol{\beta}^k \tag{14}$$

$$\boldsymbol{\xi}^{k+1} = \breve{\boldsymbol{X}}\boldsymbol{\beta}^{k+1} - \bar{\eta}_* \left\{\boldsymbol{Y} - g\left(\operatorname{prox}_{\bar{\eta}_* G}(\boldsymbol{\xi}^k + \bar{\eta}_*\boldsymbol{Y})\right)\right\}. \tag{15}$$

In the Step (iii), our goal is to ensure the algorithmic convergence of scaled GAMP iterates $\widehat{\boldsymbol{\beta}}^k = \boldsymbol{\beta}^k/\sqrt{n}$ to the estimator $\widehat{\boldsymbol{\beta}}$ as $k \to \infty$. Denote $\ell(\boldsymbol{b}) \equiv \sum_{i=1}^n \ell(\boldsymbol{b}; \boldsymbol{X}_i, Y_i)$ for any $\boldsymbol{b} \in \mathbb{R}^p$. By Taylor's theorem, we have

$$\ell(\widehat{\boldsymbol{\beta}}) = \ell(\widehat{\boldsymbol{\beta}}^k) + \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^k\right)^\top \nabla\ell(\widehat{\boldsymbol{\beta}}^k) + \frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^k\right)^\top \nabla^2\ell\left(t\widehat{\boldsymbol{\beta}} + (1-t)\widehat{\boldsymbol{\beta}}^k\right)\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^k\right),$$

for some $t \in (0,1)$. Thus, Lemma 4 implies, for some non-increasing positive function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$,

$$\ell(\widehat{\boldsymbol{\beta}}) \geq \ell(\widehat{\boldsymbol{\beta}}^k) + \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^k\right)^\top \nabla\ell(\widehat{\boldsymbol{\beta}}^k) + \frac{1}{2}n\omega\left(\max\left\{\|\widehat{\boldsymbol{\beta}}\|, \|\widehat{\boldsymbol{\beta}}^k\|\right\}\right)\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^k\|^2,$$

with probability at least $1 - c_1\exp(-c_2 n)$ where $c_1, c_2 > 0$ are some positive constants. Here, we use $\omega(\|t\widehat{\boldsymbol{\beta}} + (1-t)\widehat{\boldsymbol{\beta}}^k\|) \geq \omega(\max(\|\widehat{\boldsymbol{\beta}}\|, \|\widehat{\boldsymbol{\beta}}^k\|))$ since $\omega(\cdot)$ is non-increasing. Using optimality of the estimator $\ell(\widehat{\boldsymbol{\beta}}^k) \geq \ell(\widehat{\boldsymbol{\beta}})$ and the Cauchy-Schwarz inequality as $(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^k)^\top \nabla\ell(\widehat{\boldsymbol{\beta}}^k) \geq -\|\nabla\ell(\widehat{\boldsymbol{\beta}}^k)\|\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^k\|$ yields, with probability at least $1 - c_1\exp(-c_2 n)$,

$$\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^k\| \leq \frac{2}{\omega\left(\max\left\{\|\widehat{\boldsymbol{\beta}}\|, \|\widehat{\boldsymbol{\beta}}^k\|\right\}\right)}\left\|\frac{1}{n}\nabla\ell(\widehat{\boldsymbol{\beta}}^k)\right\| \leq \frac{2}{\omega(\|\widehat{\boldsymbol{\beta}}\|)\omega(\|\widehat{\boldsymbol{\beta}}^k\|)}\left\|\frac{1}{n}\nabla\ell(\widehat{\boldsymbol{\beta}}^k)\right\|,$$

where the last inequality follows from the fact that $0 < \omega(\cdot) < 1$ and ω is non-increasing. Next, we consider controlling $\|\nabla\ell(\widehat{\boldsymbol{\beta}}^k)\|$. We have

$$\begin{aligned}
\operatorname{prox}_{\bar{\eta}_* G}\left(\boldsymbol{\xi}^{k-1} + \bar{\eta}_*\boldsymbol{Y}\right) &= \boldsymbol{\xi}^{k-1} + \bar{\eta}_*\boldsymbol{Y} - \bar{\eta}_* g\left(\operatorname{prox}_{\bar{\eta}_* G}\left(\boldsymbol{\xi}^{k-1} + \bar{\eta}_*\boldsymbol{Y}\right)\right) \\
&= \boldsymbol{\xi}^{k-1} - \boldsymbol{\xi}^k + \breve{\boldsymbol{X}}\boldsymbol{\beta}^k \\
&= \boldsymbol{\xi}^{k-1} - \boldsymbol{\xi}^k + \boldsymbol{X}\widehat{\boldsymbol{\beta}}^k,
\end{aligned}$$

by the definition of the proximal operator and (15). Thus,

$$\begin{aligned}
\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1} &= \frac{\bar{\eta}_*}{\kappa}\breve{\boldsymbol{X}}^\top \left\{\boldsymbol{Y} - g\left(\operatorname{prox}_{\bar{\eta}_* G}(\boldsymbol{\xi}^{k-1} + \bar{\eta}_*\boldsymbol{Y})\right)\right\} \\
&= \frac{\bar{\eta}_*}{\kappa}\breve{\boldsymbol{X}}^\top \left\{\boldsymbol{Y} - g\left(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^k + \boldsymbol{\xi}^{k-1} - \boldsymbol{\xi}^k\right)\right\},
\end{aligned}$$

by (14). Using this and triangle inequalities give

$$\begin{aligned}
\left\|\nabla \ell(\widehat{\boldsymbol{\beta}}^k)\right\| &= \left\|\boldsymbol{X}^\top \left\{g(\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^k) - \boldsymbol{Y}\right\}\right\| \\
&\leq \left\|\boldsymbol{X}^\top \left\{\boldsymbol{Y} - g(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^k + \boldsymbol{\xi}^{k-1} - \boldsymbol{\xi}^k)\right\}\right\| \\
&\quad + \left\|\boldsymbol{X}^\top \left\{g(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^k + \boldsymbol{\xi}^{k-1} - \boldsymbol{\xi}^k) - g(\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^k)\right\}\right\| \\
&\leq \frac{p}{\bar{\eta}_*} \left\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}\right\| + \|\boldsymbol{X}\|_{\mathrm{op}} \left\|g(\boldsymbol{X}\widehat{\boldsymbol{\beta}}^k + \boldsymbol{\xi}^{k-1} - \boldsymbol{\xi}^k) - g(\boldsymbol{X}^\top \widehat{\boldsymbol{\beta}}^k)\right\| \\
&\leq \frac{p}{\bar{\eta}_*} \left\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}\right\| + L_g \|\boldsymbol{X}\|_{\mathrm{op}} \left\|\boldsymbol{\xi}^{k-1} - \boldsymbol{\xi}^k\right\|,
\end{aligned}$$

where $L_g = \sup_z g'(z)$. This establishes, with probability at least $1 - c_1 \exp(-c_2 n)$,

$$\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^k\| \leq c \left\{ \frac{\kappa}{\bar{\eta}_*} \left\|\widehat{\boldsymbol{\beta}}^k - \widehat{\boldsymbol{\beta}}^{k-1}\right\| + \frac{C_g}{n} \|\boldsymbol{X}\|_{\mathrm{op}} \left\|\boldsymbol{\xi}^{k-1} - \boldsymbol{\xi}^k\right\| \right\},$$

with $c = 2/\left(\omega(\|\widehat{\boldsymbol{\beta}}\|)\omega(\|\widehat{\boldsymbol{\beta}}^k\|)\right)$. Finally, Lemma 5 and the Borel-Cantelli lemma implies

$$\lim_{k\to\infty} \lim_{n\to\infty} \left\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^k\right\| =_{\mathrm{a.s.}} 0.$$

This completes the proof. □

Lemma 1 provides a statement on the convergence of the distance between 1-dimensional distributions, which is computed by averaging over the coordinates of $\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta} \in \mathbb{R}^p$. However, for statistical inference purposes, it is necessary to investigate the limiting behavior of the marginal distributions of each coordinate. To achieve this, we utilize a property of the surrogate loss function.

Lemma 2. *For any invertible matrix $\boldsymbol{L} \in \mathbb{R}^{p\times p}$ satisfying $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$, $\boldsymbol{L}^\top \widehat{\boldsymbol{\beta}}$ minimizes the surrogate loss function in (5) for the true coefficient $\boldsymbol{L}^\top \boldsymbol{\beta}$ and the covariate $\boldsymbol{L}^{-1}\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$.*

Proof of Lemma 2. Since the surrogate loss depends on $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{b} \in \mathbb{R}^p$ only through their inner product $\boldsymbol{X}^\top \boldsymbol{b}$, we have $\ell(\boldsymbol{b}; \boldsymbol{X}, Y) = \ell(\boldsymbol{L}^\top \boldsymbol{b}; \boldsymbol{L}^{-1}\boldsymbol{X}, Y)$. □

Using this in reverse, once we show Lemma A.1 for a design with identity covariance, we can rewrite the estimator corresponding to the unit covariance in $\boldsymbol{L}^\top \widehat{\boldsymbol{\beta}}$ to obtain the general covariance result.

A.3.1. Limit of Estimation and Classification Errors

As the consequence of the master theorem of GAMP (Lemma 1), we obtain the convergence limits of the mean squared error (MSE) and cosine similarity-like classification error. If we set $\psi(s,t) = (s-t)^2$, then Lemma 1 implies the MSE limit

$$\frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \xrightarrow{\mathrm{a.s.}} (\mu - 1)^2 \mathbb{E}[\bar{\beta}^2] + \sigma^2.$$

By setting $\psi(s,t) = (s-\mu t)^2$, we have a corrected MSE limit

$$\frac{1}{p}\|\widehat{\boldsymbol{\beta}} - \mu\boldsymbol{\beta}\|_2^2 \xrightarrow{\text{a.s.}} \sigma^2.$$

For classification errors, by taking the ratio of $\psi(s,t) = st$ and $\psi(s,t) = t^2$, we can obtain the convergence limit of a cosine similarity-like measure

$$\frac{\widehat{\boldsymbol{\beta}}^\top \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|^2} \xrightarrow{\text{a.s.}} \mu.$$

A.4. GLM extension of the SLOE estimator

This section provides an extension of the SLOE estimator [YYMD21] from logistic regression to the GLM. First, it considers another representation of the state evolution parameters by using *corrupted signal strength* $\gamma_c^2 = \lim_{n\to\infty} \mathrm{Var}(\boldsymbol{X}_1(n)^\top \widehat{\boldsymbol{\beta}}(n))$ instead of γ^2:

$$\begin{cases} \kappa^2\sigma^2 & = \eta^2 \mathbb{E}_{(Q_1', Q_2', \bar{Y})}\left[\left(\bar{Y} - g\left(\mathrm{prox}_{\eta G}(Q_2')\right)\right)^2\right], \\ 0 & = \mathbb{E}_{(Q_1', Q_2', \bar{Y})}\left[Q_1'\left(\bar{Y} - g\left(\mathrm{prox}_{\eta G}(Q_2')\right)\right)\right], \\ 1-\kappa & = \mathbb{E}_{(Q_1', Q_2', \bar{Y})}\left[\left(1 + \eta g'\left(\mathrm{prox}_{\eta G}(Q_2')\right)\right)^{-1}\right], \end{cases}$$

where

$$\begin{pmatrix} Q_1' \\ Q_2' \end{pmatrix} \sim \mathcal{N}_2\left(0, \begin{bmatrix} \mu^{-2}(\gamma_c^2 - \kappa\sigma^2) & -\mu^{-1}(\gamma_c^2 - \kappa\sigma^2) \\ -\mu^{-1}(\gamma_c^2 - \kappa\sigma^2) & \gamma_c^2 \end{bmatrix}\right).$$

Then, since $\boldsymbol{X}_1^\top \widehat{\boldsymbol{\beta}}$ is computable unlike $\boldsymbol{X}_1^\top \boldsymbol{\beta}$, we can construct the SLOE-like estimator for γ_c^2 as follows.

$$\begin{aligned} \widehat{\gamma}_c^2 &= \frac{1}{n}\sum_{i=1}^n S_i^2 - \left(\frac{1}{n}\sum_{i=1}^n S_i\right)^2, \\ S_i &= \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}} + \frac{U_i}{1 + g'(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})}(Y_i - g(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})), \\ U_i &= -(\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{X}^\top)_{ii}, \end{aligned}$$

where $\boldsymbol{D} = \mathrm{diag}(g'(\boldsymbol{X}_1^\top \widehat{\boldsymbol{\beta}}), \ldots, g'(\boldsymbol{X}_n^\top \widehat{\boldsymbol{\beta}})) \in \mathbb{R}^{n\times n}$. In fact, this is a consistent estimator of γ_c^2.

Proposition 5. *Suppose that $\widehat{\boldsymbol{\beta}}$ exists and $\kappa \in (0,1)$ is fixed. Under Assumption 1 (A1)–(A2), $\widehat{\gamma}_c^2 \overset{\text{a.s.}}{\to} \gamma_c^2$ as $n \to \infty$.*

This follows from the direct application of the original proof of Proposition 2 in [YYMD21] because of the form of the surrogate loss and the monotonicity of the link function $g(\cdot)$.

Unfortunately, the estimator $\widehat{\gamma}_c^2$ is unstable when p/n is to 1. This property inherits from two components: $\widehat{\boldsymbol{\beta}}$ and $(\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X})^{-1}$ in U_i. First, for large p/n, $\widehat{\boldsymbol{\beta}}$ is likely to diverge as evident from [CS20]. This phenomenon is observed in many estimators. Second, the upper bound of $|\widehat{\gamma}_c^2 - \gamma_c^2|$ depends on a constant $1/\lambda_{\min}(\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X})$. This constant diverges as $\kappa \uparrow 1$.

A.5. Regularized Estimator

In this section, we consider the following regularized estimator:

$$\widehat{\boldsymbol{\beta}}_\lambda \in \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \sum_{i=1}^n \ell(\boldsymbol{b}; \boldsymbol{X}_i, Y_i) + \lambda \sum_{j=1}^p J(b_j) \right\},$$

where $J : \mathbb{R} \to \mathbb{R}$ is some regularization function and $\lambda > 0$ is a tuning parameter. It is important to consider the regularized estimator $\widehat{\boldsymbol{\beta}}_\lambda$, because in some cases such as logistic regression with $n < 2p$, the unregularized estimator $\widehat{\boldsymbol{\beta}}$ does not exist (see, for example, [CS20]). The regularization imposes constraints on the estimator within specific regions around the origin. Consequently, the issue of non-existence can be mitigated by employing a suitable regularized estimator. Note that even when we employ the regularized estimator $\widehat{\boldsymbol{\beta}}_\lambda$, the method of estimating γ^2 and σ_e^2 is not affected by the method of estimating $\boldsymbol{\beta}$.

We display the system of nonlinear equations which characterizes the state evolution parameters in L_2 penalized cases. If we set $J(t) = t^2$, we have

$$\begin{cases} \kappa^2\sigma^2 & = \eta^2 \mathbb{E}_{(Q_1,Q_2,\bar{Y})}\left[\left(\bar{Y} - g\left(L\right)\right)^2 \right], \\ 2\gamma^2\lambda\mu & = \mathbb{E}_{(Q_1,Q_2,\bar{Y})}\left[Z\left(\bar{Y} - g\left(L\right)\right)\right], \\ 1-\kappa+2\lambda\eta & = \mathbb{E}_{(Q_1,Q_2,\bar{Y})}\left[\left(1+\eta g'\left(L\right)\right)^{-1}\right], \end{cases} \tag{16}$$

where $L = \operatorname{prox}_{\eta G}(\mu Z + \sqrt{\kappa}\sigma Q_2 + \eta\bar{Y})$, $Z = \gamma Q_1$, $(Q_1, Q_2) \sim \mathcal{N}_2(0, \boldsymbol{I}_2)$ and $\bar{Y} = h(Z, \bar{\varepsilon})$. We can see that (16) admits the case $n < p$ since the left-hand side of the last equation can be positive with sufficiently large $\lambda > 0$ while the right-hand side is always positive.

Since regularized estimators are biased, the distributional characterization of the limit of $\widehat{\boldsymbol{\beta}}_\lambda$ is somewhat different from the unregularized case. Actually, we have an extension of Lemma 1 as

$$\frac{1}{p}\sum_{j=1}^p \psi\left(\sqrt{n}\widehat{\beta}_j, \sqrt{n}\beta_j\right) \xrightarrow{\text{a.s.}} \mathbb{E}[\psi\left(\operatorname{prox}_{\eta J}(\mu\bar{\beta} + \sigma G), \bar{\beta}\right)],$$

under the settings of Lemma 1. For the unregularized case $J \equiv$ const., this is reduced to Lemma 1. Suppose that $J(\cdot)$ is differentiable and strongly convex, and $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$ here. Using the fact that $\operatorname{prox}_{\eta J}(x) = x - \eta J'(\operatorname{prox}_{\eta J}(x))$ for $x \in \mathbb{R}$ and $\eta > 0$ by the definition of the proximal operator, we finally obtain

$$\sqrt{n}\frac{\widehat{\beta}_j^{(\mathrm{d})} - \mu\beta_j}{\sigma} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\widehat{\boldsymbol{\beta}}^{(\mathrm{d})}$ is a debiased estimator

$$\widehat{\boldsymbol{\beta}}^{(\mathrm{d})} = \widehat{\boldsymbol{\beta}}_\lambda + \eta J'(\widehat{\boldsymbol{\beta}}_\lambda).$$

A.6. GLM Form with $h(Z, \bar{\varepsilon})$

In this section, we discuss the other form of GLM used in (A3) in Assumption 1 as follows:

$$Y_i = h\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}, \varepsilon_i\right), \tag{17}$$

where $h : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is a deterministic function determined by the distribution of $Y \mid \boldsymbol{X}$, and $\varepsilon_i \in \mathbb{R}$ is an error variable *independent* of $\boldsymbol{X}$. While the GLM 1 only specifies the conditional mean, (17) completely determines the distributional behavior of Y depending on $\boldsymbol{X}$. The major difference from the additive model (7) used in our estimation is that the random variable ε_i is independent from $\boldsymbol{X}_i^\top$.

This model (17) is flexible in design and allows for an intuitive representation of GLMs. We see several examples of $h(Z, \bar{\varepsilon})$ introduced in (6).

• *Bernoulli case (binary choice model).* When we use a model $\bar{Y}|Z \sim \mathrm{Ber}(g(Z))$ with some $g : \mathbb{R} \to [0, 1]$, the inverse transformation method yields

$$\bar{Y} = 1\{g(Z) \le \bar{\varepsilon}\}, \quad \bar{\varepsilon} \sim \mathrm{Unif}[0, 1].$$

• *Exponential case.* If $\bar{Y}|Z \sim \mathrm{Exp}(g(Z))$ with some $g : \mathbb{R} \to (0, \infty)$, the inverse transformation method yields

$$\bar{Y} = -\frac{1}{g(Z)} \log(\bar{\varepsilon}), \quad \bar{\varepsilon} \sim \mathrm{Unif}[0, 1].$$

• *Poisson case.* If $\bar{Y}|Z \sim \mathrm{Pois}(g(Z))$ with some $g : \mathbb{R} \to (0, \infty)$, we have

$$\bar{Y} = \min\left\{k \in \mathbb{N} \cup \{0\} \,\middle|\, \sum_{l=1}^{k+1} \bar{\varepsilon}_l > g(Z)\right\}, \quad \bar{\varepsilon}_l \overset{\mathrm{iid}}{\sim} \mathrm{Exp}(1).$$

• *Gaussian case.* If $\bar{Y}|Z \sim \mathcal{N}(g(Z), \sigma_{\bar{\varepsilon}}^2)$ with some $g : \mathbb{R} \to \mathbb{R}$, we have

$$\bar{Y} = g(Z) + \bar{\varepsilon}, \quad \bar{\varepsilon} \sim \mathcal{N}(0, \sigma_{\bar{\varepsilon}}^2).$$

As evident from the aforementioned examples, in the case of modeling using a one-parameter distribution, the distribution of $\bar{\varepsilon}$ can be fully determined without any additional parameters.

A.7. Construction of CI in Real Data Application

This section specifies the construction of the classical and proposed CI in Section 5.3. To begin with, we compute the MLE $\widehat{\boldsymbol{\beta}}$ for the Poisson regression model. A classical MLE theory implies, for any $\boldsymbol{x} \in \mathbb{R}^p$,

$$\sqrt{n}(\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{x}^\top \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{x}^\top \mathcal{I}_{\boldsymbol{\beta}}^{-1} \boldsymbol{x}),$$

as $n \to \infty$ with fixed p. Using this, we construct the classical CI of each $\boldsymbol{X}_i^\top \boldsymbol{\beta}$ with a preassigned level $(1-\alpha)$ as

$$\left[\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}} - z_{(1-\alpha/2)} \sqrt{\frac{\boldsymbol{X}_i^\top \widehat{\mathcal{I}}(\widehat{\boldsymbol{\beta}}) \boldsymbol{X}_i}{n}}, \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}} + z_{(1-\alpha/2)} \sqrt{\frac{\boldsymbol{X}_i^\top \widehat{\mathcal{I}}(\widehat{\boldsymbol{\beta}}) \boldsymbol{X}_i}{n}} \right].$$

We have proposed a CI for $\beta_j, j = 1, \ldots, p$ in (13). Using a similar technique, we can construct a valid CI for $\boldsymbol{X}_i^\top \boldsymbol{\beta}, i = 1, \ldots, n$ in proportionally high dimensions. Actually, we obtain

Proposition 6. *Under the setting of Theorem 2, we have, for each $i = 1, \ldots, n$,*

$$\frac{\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}} + \eta \ell'(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}) - \mu_Z \boldsymbol{X}_i^\top \boldsymbol{\beta}}{\sigma_Z} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n, p(n) \to \infty$ with $p(n)/n \to \kappa$. Here, $\ell'(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}) = g(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}) - Y_i$.

This proposition yields the asymptotically level $(1-\alpha)$ confidence interval of $\boldsymbol{X}_i^\top \boldsymbol{\beta}$,

$$\frac{1}{\mu_Z} \left[\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}} + \eta \ell'(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}) - \sigma_Z z_{(1-\alpha/2)}, \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}} + \eta \ell'(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}) + \sigma_Z z_{(1-\alpha/2)} \right].$$

Proof of Proposition 6. Since $\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}} = (\boldsymbol{L}^{-1} \boldsymbol{X}_i)^\top (\boldsymbol{L}^\top \widehat{\boldsymbol{\beta}})$ with $\boldsymbol{L}^{-1} \boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$, we can repeat the arguments in Proposition 3 by replacing $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$ with $(\boldsymbol{X}_1^\top \boldsymbol{\beta}, \ldots, \boldsymbol{X}_n^\top \boldsymbol{\beta})^\top$ and $(\boldsymbol{X}_1^\top \widehat{\boldsymbol{\beta}} + \eta \ell'(\boldsymbol{X}_1^\top \widehat{\boldsymbol{\beta}}), \ldots, \boldsymbol{X}_n^\top \widehat{\boldsymbol{\beta}} + \eta \ell'(\boldsymbol{X}_n^\top \widehat{\boldsymbol{\beta}}))$, respectively. □

A.8. Sufficient Condition for Assumption 2

We discuss a sufficient condition for Assumption 2 (A5). We require the monotonicity below to identify the SS parameter by using the method provided in Section 3.2.3.

Lemma 3. *Let $\mathbb{E}[g(Z_\varsigma)] < \infty$ for $\varsigma > 0$ and g_0 be strictly monotone on $\mathbb{R}_+$. Then, the map $\varsigma \mapsto \mathbb{E}[g(Z_\varsigma)]$ is strictly monotonic in $\varsigma > 0$.*

This ensures the uniqueness of the estimator $\widehat{\gamma}^2$. One of the sufficient conditions is the following.

Assumption 3. *The odd part $g_0(x) := (g(x) + g(-x))/2$ of the inverse link function $g(\cdot)$ is strictly monotonic for $\mathbb{R}_+$.*

Intuitively, this implies that the form of $g(\cdot)$ is not point-symmetric around point $(0, g(0))$, which is a generalization of a non-odd function. This precludes logistic regression, but many other models satisfy the condition. However, a later study [CLM24] proposes a method that utilizes information from $\boldsymbol{X}$ to handle logistic regression using a moment-based technique, specifically when $n > p$.

Appendix B: Proofs of Main Results

Proof of Proposition 2. Since the surrogate loss has an equivalent form of the negative log-likelihood function for logistic regression, it can be derived according to the discussion of [FVRS22] in Section 4.7. Note that the construction of $h(\cdot,\cdot)$ and the inverse link function $g(\cdot)$ are generalized in our case. □

Proof of Proposition 3. Lemma 2 implies that, for any $j = 1, \dots, p$,

$$\frac{\widehat{\beta}_j - \mu\beta_j}{\sigma/\tau_j} = \frac{\widehat{\theta}_j - \mu\theta_j}{\sigma},$$

where we define

$$\boldsymbol{\theta} = \boldsymbol{L}^\top \boldsymbol{\beta}, \quad \widehat{\boldsymbol{\theta}} = \boldsymbol{L}^\top \widehat{\boldsymbol{\beta}}, \tag{18}$$

by a Cholesky factorization $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$. We have $\widehat{\theta}_j = \tau_j\widehat{\beta}_j$ and $\theta_j = \tau_j\beta_j$ by the rearrangement of indices. Here, define

$$\mu_n = \frac{\widehat{\boldsymbol{\theta}}^\top \boldsymbol{\theta}}{\|\boldsymbol{\theta}\|^2}, \qquad \text{and} \qquad \sigma_n^2 = \frac{1}{\kappa}\|\widehat{\boldsymbol{\theta}} - \mu_n\boldsymbol{\theta}\|^2. \tag{19}$$

Then, we have

$$\sqrt{n}\frac{\widehat{\theta}_j - \mu\theta_j}{\sigma} = \sqrt{n}\frac{\widehat{\theta}_j - \mu_n\theta_j}{\sigma_n}\frac{\sigma_n}{\sigma} + \sqrt{n}\frac{(\mu_n - \mu)\theta_j}{\sigma}.$$

Lemma 7 gives us

$$\frac{\widehat{\boldsymbol{\theta}} - \mu_n\boldsymbol{\theta}}{\sigma_n} \overset{\mathrm{d}}{=} \frac{\boldsymbol{P}_{\boldsymbol{\theta}}^\perp \boldsymbol{Z}}{\|\boldsymbol{P}_{\boldsymbol{\theta}}^\perp \boldsymbol{Z}\|}, \tag{20}$$

where $\boldsymbol{Z} = (Z_1, \dots, Z_p) \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Triangle inequalities yield that

$$\frac{\|\boldsymbol{Z}\|}{\sqrt{p}} - \frac{|\boldsymbol{\theta}^\top \boldsymbol{Z}|}{\sqrt{p}\,\|\boldsymbol{\theta}\|} \le \frac{\|\boldsymbol{P}_{\boldsymbol{\theta}}^\perp \boldsymbol{Z}\|}{\sqrt{p}} \le \frac{\|\boldsymbol{Z}\|}{\sqrt{p}} + \frac{|\boldsymbol{\theta}^\top \boldsymbol{Z}|}{\sqrt{p}\,\|\boldsymbol{\theta}\|}.$$

Since $|\boldsymbol{\theta}^\top \boldsymbol{Z}|/(\sqrt{p}\,\|\boldsymbol{\theta}\|) \xrightarrow{\text{a.s.}} 0$ and $\|\boldsymbol{Z}\|/\sqrt{p} \xrightarrow{\text{a.s.}} 1$, we have $\|\boldsymbol{P}_{\boldsymbol{\theta}}^\perp \boldsymbol{Z}\|/\sqrt{p} \xrightarrow{\text{a.s.}} 1$. Then, this fact and (20) imply that

$$\sqrt{n}\frac{\widehat{\theta}_j - \mu_n\theta_j}{\sigma_n} \overset{\mathrm{d}}{=} \frac{1}{\sqrt{\kappa}}\sigma_j Q + o_p(1), \qquad \sigma_j^2 = 1 - \frac{\theta_j^2}{\|\boldsymbol{\theta}\|^2},$$

where $Q \sim \mathcal{N}(0,1)$. Here we use the fact that the covariance matrix of $\boldsymbol{P}_{\boldsymbol{\theta}}^\perp \boldsymbol{Z}$ is $\boldsymbol{P}_{\boldsymbol{\theta}}^\perp \boldsymbol{P}_{\boldsymbol{\theta}}^\perp = \boldsymbol{I}_p - \boldsymbol{\theta}\boldsymbol{\theta}^\top/\|\boldsymbol{\theta}\|^2$. Thus, the facts that $\mu_n \xrightarrow{\text{a.s.}} \mu$ and $\sigma_n^2 \xrightarrow{\text{a.s.}} \sigma^2$ by Lemma 8 conclude the proof. □

Proof of Lemma 3. We write the cumulative distribution and density functions of the standard normal distribution by $\Phi : \mathbb{R} \to \mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}$, respectively. Note that a density function of Gaussian distribution with mean zero is an even function. Since a product of even functions is even and a product of an even function and an odd function is odd, we have

$$\mathbb{E}[g(Z_\varsigma)] = \frac{1}{\varsigma}\int_{-\infty}^{\infty} g(x)\phi\left(\frac{x}{\varsigma}\right) dx = \int_{-\infty}^{\infty} g(\varsigma y)\phi\left(y\right) dy = 2\int_{0}^{\infty} g_0(\varsigma y)\phi\left(y\right) dy,$$

where the second identity is from a change of variables $y = x/\varsigma$. Thus, For any $\varsigma_1 > \varsigma_0 > 0$, we can say that

$$\mathbb{E}[g(Z_{\varsigma_1})] - \mathbb{E}[g(Z_{\varsigma_0})] = 2\int_{0}^{\infty} \left\{g_0(\varsigma_1 y) - g_0(\varsigma_0 y)\right\}\phi\left(y\right) dy.$$

is strictly positive or negative by $\phi(\cdot) > 0$. Therefore, if the even part $g_0(\cdot)$ of $g(\cdot)$ is monotone on $\mathbb{R}_+$, then $\mathbb{E}[g(Z_\varsigma)]$ is monotone in $\varsigma > 0$. □

Proof of Theorem 1. **Case of** (12). The law of large numbers yields

$$\sup_{\varsigma^2 > 0} \left|\tilde{\Psi}_n(\varsigma^2) - \tilde{\Psi}(\varsigma^2)\right| = \left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}[\bar{Y}]\right| \xrightarrow{\text{a.s.}} 0, \tag{21}$$

Thus, (21), Assumption (A5), and $\tilde{\Psi}(\gamma) = 0$ imply $\widehat{\gamma}^2 \xrightarrow{\text{a.s.}} \gamma^2$ by Theorem 5.9 in [vdV00].

Case of (10) By the definition, we have

$$\sup_{\varsigma,\varsigma_e} \|\Psi_n(\varsigma,\varsigma_e) - \Psi(\varsigma,\varsigma_e)\| = \left\|\left(n^{-1}\sum_{i=1}^{n} Y_i^2 - \mathbb{E}[\bar{Y}^2], n^{-1}\sum_{i=1}^{n} Y_i^4 - \mathbb{E}[\bar{Y}^4]\right)\right\| \xrightarrow{\text{a.s.}} 0,$$

where the convergence follows from the law of large numbers. This uniform convergence, Assumption (A5), and $\Psi(\gamma) = \mathbf{0}$ imply $(\widehat{\gamma}^2, \widehat{\sigma}_e^2) \xrightarrow{\text{a.s.}} (\gamma^2, \sigma_e^2)$ by Theorem 5.9 in [vdV00]. □

Proof of Theorem 2. By Proposition 4, $\mathcal{CI}_{\alpha,j}$ is clearly asymptotically level $(1-\alpha)$ confidence interval, since we have

$$\begin{aligned}
&\Pr\left(\frac{\widehat{\beta}_j}{\widehat{\mu}} + z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}\widehat{\mu}\widehat{\tau}_j} \le \beta_j \le \frac{\widehat{\beta}_j}{\widehat{\mu}} - z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}\widehat{\mu}\widehat{\tau}_j}\right) \\
&= \Pr\left(z_{\alpha/2} \le \frac{\sqrt{n}(\widehat{\beta}_j - \widehat{\mu}\beta_j)}{\widehat{\sigma}/\widehat{\tau}_j} \le z_{(1-\alpha/2)}\right) \\
&\to (1-\alpha).
\end{aligned}$$

□

Proof of Proposition 4. We have

$$\begin{aligned}\sqrt{n}\widehat{\tau}_j\frac{\widehat{\beta}_j - \widehat{\mu}\beta_j}{\widehat{\sigma}} &= \sqrt{n}\widehat{\tau}_j\frac{\widehat{\beta}_j - \mu\beta_j}{\sigma}\frac{\sigma}{\widehat{\sigma}} - \sqrt{n}\widehat{\tau}_j\frac{(\mu - \widehat{\mu})\beta_j}{\widehat{\sigma}} \\ &= \sqrt{n}\widehat{\tau}_j\frac{\widehat{\beta}_j - \mu\beta_j}{\sigma} - \sqrt{n}\widehat{\tau}_j\frac{\widehat{\beta}_j - \mu\beta_j}{\sigma}\frac{\widehat{\sigma} - \sigma}{\widehat{\sigma}} - \sqrt{n}\widehat{\tau}_j\frac{(\mu - \widehat{\mu})\beta_j}{\widehat{\sigma}} \\ &= \sqrt{n}\tau_j\frac{\widehat{\beta}_j - \mu\beta_j}{\sigma} - o_p(1),\end{aligned}$$

where the last identity follows from Lemma 6, $\sqrt{n}\tau_j\beta_j = O(1)$, and the assumption (A1). Finally, Proposition 3 concludes the proof. □

Appendix C: Technical Lemmas

For a matrix $A \in \mathbb{R}^{m\times m}$, $A \succeq 0$ is defined to mean that A is positive semidefinite.

Lemma 4. *Suppose that $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Under the assumption (A2), for some constant ϵ_0 such that $0 \le \epsilon \le \epsilon_0$,*

$$\frac{1}{n}\sum_{i=1}^{n}\nabla^2\ell(\boldsymbol{b};\boldsymbol{X}_i,Y_i) \succeq \left(\inf_{z:|z|\le\frac{3\|\boldsymbol{b}\|}{\sqrt{\epsilon}}} g'(z)\right)\left(\sqrt{1-\epsilon} - \sqrt{\kappa} - 2\sqrt{\frac{H(\epsilon)}{1-\epsilon}}\right)^2\boldsymbol{I}_p$$

with $H(\epsilon) = -\epsilon\log\epsilon - (1-\epsilon)\log(1-\epsilon)$ holds for any $\boldsymbol{b} \in \mathbb{R}^p$ with probability at least $1 - 2\exp(-nH(\epsilon)) - 2\exp(-n/2)$.

Proof of Lemma 4. Since the proof of Lemma 3 in [SCC19] only uses the specific structure $\nabla^2\ell(\boldsymbol{b};\boldsymbol{X}_i,Y_i) = \rho''(\boldsymbol{X}_i^\top\boldsymbol{b})\boldsymbol{X}_i\boldsymbol{X}_i^\top$ with $\rho'(t) = 1/(1+\exp(-t))$, we can immediately extend it to our general structure $\nabla^2\ell(\boldsymbol{b};\boldsymbol{X}_i,Y_i) = g'(\boldsymbol{X}_i^\top\boldsymbol{b})\boldsymbol{X}_i\boldsymbol{X}_i^\top$. □

Remark 2. *This lemma implies that, for sufficiently small $\epsilon > 0$,*

$$\frac{1}{n}\sum_{i=1}^{n}\nabla^2\ell(\boldsymbol{b};\boldsymbol{X}_i,Y_i) \succeq \omega(\|\boldsymbol{b}\|)\boldsymbol{I}_p,$$

for some non-increasing positive function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ independent of n.

Lemma 5. *Consider the setting of Lemma 1. For the generalized approximate message passing recursion (14)-(15), we have the Cauchy property of the recursion:*

$$\lim_{k\to\infty}\lim_{n\to\infty}\|\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^k\|^2 =_{\text{a.s.}} 0,$$
$$\lim_{k\to\infty}\lim_{n\to\infty}\|\boldsymbol{\xi}^{k+1} - \boldsymbol{\xi}^k\|^2 =_{\text{a.s.}} 0.$$

Proof of Lemma 5. For the stationary version of the GAMP recursion, we have

$$\bar{\eta}_* \left\{ \boldsymbol{Y} - g\left(\operatorname{prox}_{\bar{\eta}_* G}(\boldsymbol{\xi}^k + \bar{\eta}_* \boldsymbol{Y}) \right) \right\} = \operatorname{prox}_{\bar{\eta}_* \ell}(\boldsymbol{\xi}^k) - \boldsymbol{\xi}^k,$$

where $\operatorname{prox}_{\bar{\eta}_* \ell}(\cdot)$ depends on the first input of the surrogate loss function ℓ, by the fact that $\operatorname{prox}_{\eta G}(x) = x - g(\operatorname{prox}_{\eta G}(x))$ for any $x \in \mathbb{R}, \eta > 0$ by the definition of the proximal operator. Thus, taking $\Psi(z; b) = z - \operatorname{prox}_{b\ell}(z)$, we can straightforwardly repeat the argument of Lemma 6.9 in [DM16] and complete the proof. □

Recall that $\widehat{\eta}$ is a solution of the SE system as defined in Section 3.3.

Lemma 6. *Under the settings in Proposition 4, we have*

$$\widehat{\eta} \xrightarrow{p} \eta, \quad \widehat{\mu} \xrightarrow{p} \mu, \quad \widehat{\sigma}^2 \xrightarrow{p} \sigma^2,$$

as $n, p(n) \to \infty$ with $p(n)/n \to \kappa$.

Proof of Lemma 6. In this proof, we denote L_f as the Lipschitz constant of a function $f(\cdot)$. To emphasize the dependence on the signal strength, we denote $\eta(\widehat{\gamma}) \equiv \widehat{\eta}, \mu(\widehat{\gamma}) \equiv \widehat{\mu}$, $\sigma(\widehat{\gamma}) \equiv \widehat{\sigma}$, $\eta(\gamma) \equiv \eta$, $\mu(\gamma) \equiv \mu$, and $\sigma(\gamma) \equiv \sigma$. To begin with, note that, as discussed in Section 4.4 in [FVRS22], the solutions to the system of nonlinear equations (6) can be rewritten as fixed points of the following recursions:

$$\begin{aligned}
\eta_{k+1}(\gamma) &= \kappa \eta_k(\gamma) \left(1 - \mathbb{E}_{(Q_1,Q_2,U)} \left[\frac{1}{1 + \eta_k(\gamma) g'\left(d_k(\gamma)\right)} \right] \right)^{-1} \\
\mu_{k+1}(\gamma) &= \frac{\eta_{k+1}(\gamma)}{\gamma^2} \mathbb{E}_{(Q_1,Q_2,U)} \left[\gamma Q_1 \left\{ h(\gamma Q_1, U) - g\left(d_k(\gamma)\right) \right\} \right] \\
\sigma^2_{k+1}(\gamma) &= \frac{\eta^2_{k+1}(\gamma)}{\kappa^2} \mathbb{E}_{(Q_1,Q_2,U)} \left[\left\{ h(\gamma Q_1, U) - g\left(d_k(\gamma)\right) \right\}^2 \right],
\end{aligned}$$

where

$$\begin{aligned}
&d_k(\gamma) = \operatorname{prox}_{\eta_k(\gamma) G}(\mu_k(\gamma)\gamma Q_1 + \sqrt{\kappa}\sigma_k(\gamma) Q_2 + \eta_k(\gamma) h(\gamma Q_1, U)), \\
&(Q_1, Q_2) \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{I}_2), \qquad U \perp\!\!\!\perp (Q_1, Q_2),
\end{aligned}$$

and $h(\gamma Q_1, U)$ is designed to have the same distribution as Y.

Step 1. In this step, we inductively show that in each step $k \in \mathbb{N}$, $(\eta_k(\widehat{\gamma}), \mu_k(\widehat{\gamma}), \sigma_k(\widehat{\gamma}))$ of the recursion converges to $(\eta_k(\gamma), \mu_k(\gamma), \sigma_k(\gamma))$ using the fact that $\widehat{\gamma}^2 \xrightarrow{\text{a.s.}} \gamma^2$ by Theorem 1. Let $(\eta_0, \mu_0, \sigma_0)$ be a given triplet of the initializers, and assume $\eta_0 > 0$ and

$$\max\{|\eta_{k-1}(\widehat{\gamma}) - \eta_{k-1}(\gamma)|, |\mu_{k-1}(\widehat{\gamma}) - \mu_{k-1}(\gamma)|, |\sigma_{k-1}(\widehat{\gamma}) - \sigma_{k-1}(\gamma)|\} = o_p(1)$$

.

• **Bound for** d_0. We have

$$\begin{aligned}
|d_0(\widehat{\gamma}) - d_0(\gamma)| &= |\mathrm{prox}_{\eta_0 G}(\mu_0 \widehat{\gamma} Q_1 + \sqrt{\kappa}\sigma_0 Q_2 + \eta_0 h(\widehat{\gamma} Q_1, U)) \\
&\qquad - \mathrm{prox}_{\eta_0 G}(\mu_0 \gamma Q_1 + \sqrt{\kappa}\sigma_0 Q_2 + \eta_0 h(\gamma Q_1, U))| \\
&\leq |\mu_0(\widehat{\gamma} - \gamma) Q_1 + \eta_0 h(\widehat{\gamma} Q_1, U) - \eta_0 h(\gamma Q_1, U)| \\
&\leq ((\mu_0 + L_h \eta_0) Q_1) |\widehat{\gamma} - \gamma| := L_{d_0} |\widehat{\gamma} - \gamma|,
\end{aligned}$$

where the first inequality follows from the Lipschitz continuous of the proximal operator by Lemma 9 (i), and the last inequality is from the Lipschitz condition on $h(\cdot,\cdot)$ and the triangle inequality. Thus, $|d_0(\widehat{\gamma}) - d_0(\gamma)| = o_p(1)$

• **Bound for** η_1. We use the fact that $\left|\frac{1}{1-a} - \frac{1}{1-b}\right| \leq \frac{1}{(1-a)(1-b)} |a-b|$ for $0 < a, b < 1$, and $\left|\frac{1}{1+a} - \frac{1}{1+b}\right| \leq |a-b|$ for $0 < a, b$. Define a constant

$$C_{\eta_1} = \left(1 - \mathbb{E}\left[\frac{1}{1 + \eta_0 g'(d_0(\widehat{\gamma}))}\right]\right)^{-1} \left(1 - \mathbb{E}\left[\frac{1}{1 + \eta_0 g'(d_0(\gamma))}\right]\right)^{-1} > 0.$$

Since $1 + \eta_0 g'(\cdot) > 1$ by the monotonically increasing property of $g(\cdot)$, we have

$$\begin{aligned}
|\eta_1(\widehat{\gamma}) - \eta_1(\gamma)| &\leq \kappa \eta_0 C_{\eta_1} \mathbb{E}\left[\left|\frac{1}{1 + \eta_0 g'\left(d_0(\widehat{\gamma})\right)} - \frac{1}{1 + \eta_0 g'\left(d_0(\gamma)\right)}\right|\right] \\
&\leq \kappa \eta_0^2 C_{\eta_1} \mathbb{E}\left[|g'(d_0(\widehat{\gamma})) - g'(d_0(\gamma))|\right] \\
&\leq \kappa \eta_0^2 C_{\eta_1} L_{g'} L_{d_0} |\widehat{\gamma} - \gamma|,
\end{aligned}$$

where the last inequality uses the $L_{g'}$-smoothness of the inverse link function $g(\cdot)$ and the Lipschitz continuity of $d_0(\gamma)$. Then, $|\eta_1(\widehat{\gamma}) - \eta_1(\gamma)| = o_p(1)$.

• **Bound for** μ_1. By the triangle inequality, we have

$$\begin{aligned}
&|\mu_1(\widehat{\gamma}) - \mu(\gamma)| \\
&\leq \left|\left(\frac{\eta_1(\widehat{\gamma})}{\widehat{\gamma}} - \frac{\eta_1(\gamma)}{\gamma}\right) \mathbb{E}\left[Q_1 \left\{h(\widehat{\gamma} Q_1, U) - g(d_0(\widehat{\gamma}))\right\}\right]\right| \\
&\quad + \left|\frac{\eta_1(\gamma)}{\gamma} \mathbb{E}\left[Q_1 \left\{h(\widehat{\gamma} Q_1, U) - h(\gamma Q_1, U) - g(d_0(\widehat{\gamma})) + g(d_0(\gamma))\right\}\right]\right|.
\end{aligned}$$

Using the fact that $\frac{c}{a+b} = \frac{c}{a} - \frac{cb}{a(a+b)}$ for any $a \neq 0, b \neq -a, c \in \mathbb{R}$, we have

$$\frac{\eta_1(\widehat{\gamma})}{\widehat{\gamma}} - \frac{\eta_1(\gamma)}{\gamma} = \frac{\eta_1(\widehat{\gamma})}{\gamma} - \frac{\eta_1(\widehat{\gamma})\,(\widehat{\gamma} - \gamma)}{\gamma \widehat{\gamma}} - \frac{\eta_1(\gamma)}{\gamma} = \frac{\eta_1(\widehat{\gamma}) - \eta_1(\gamma)}{\gamma} - o_p(1) = o_p(1).$$

Thus, by the Lipschitz continuity of $h(\cdot,\cdot)$, $g(\cdot)$, and d_0, we have

$$|\mu_1(\widehat{\gamma}) - \mu_1(\gamma)| \leq \left|\frac{\eta_1(\gamma)}{\gamma} (\widehat{\gamma} - \gamma) \mathbb{E}\left[Q_1 (L_h Q_1 - L_g L_{d_0})\right]\right| + o_p(1) = o_p(1).$$

- **Bound for** σ_1. By the triangle inequality and Jensen's inequality, we have

$$
\begin{aligned}
&\left|\sigma_1^2(\widehat{\gamma}) - \sigma_1^2(\gamma)\right| \\
&\leq \frac{1}{\kappa^2}\left|\left(\eta_1^2(\widehat{\gamma}) - \eta_1^2(\gamma)\right)\mathbb{E}\left[\left\{h(\widehat{\gamma}Q_1, U) - g\left(d_0(\widehat{\gamma})\right)\right\}^2\right]\right| \\
&\quad + \frac{\eta_1^2(\gamma)}{\kappa^2}\mathbb{E}\left|\left\{h(\widehat{\gamma}Q_1, U) - g\left(d_0(\widehat{\gamma})\right)\right\}^2 - \left\{h(\gamma Q_1, U) - g\left(d_0(\gamma)\right)\right\}^2\right| \\
&\leq \frac{1}{\kappa^2}\left|\widehat{\gamma} - \gamma\right|\left(\eta_1(\widehat{\gamma}) + \eta_1(\gamma)\right)\mathbb{E}\left[\left\{h(\widehat{\gamma}Q_1, U) - g\left(d_0(\widehat{\gamma})\right)\right\}^2\right] \\
&\quad + \frac{\eta_1^2(\gamma)}{\kappa^2}\mathbb{E}\left[\left|\left\{h(\widehat{\gamma}Q_1, U) - g\left(d_0(\widehat{\gamma})\right)\right\} - \left\{h(\gamma Q_1, U) - g\left(d_0(\gamma)\right)\right\}\right|\right. \\
&\quad \left.\cdot\left\{h(\widehat{\gamma}Q_1, U) - g\left(d_0(\widehat{\gamma})\right) + h(\gamma Q_1, U) - g\left(d_0(\gamma)\right)\right\}\right] \\
&\leq o_p(1) + \frac{\eta_1^2(\gamma)}{\kappa^2}\left|\widehat{\gamma} - \gamma\right| \\
&\quad \times \mathbb{E}\left|\left(L_h Q_1 - L_g L_{d_0}\right)\left(h(\widehat{\gamma}Q_1, U) - g\left(d_0(\widehat{\gamma})\right) + h(\gamma Q_1, U) - g\left(d_0(\gamma)\right)\right)\right| \\
&= o_p(1).
\end{aligned}
$$

Also, since $|a - b| \leq \sqrt{|a^2 - b^2|}$ for $a, b > 0$, we have $|\sigma_1(\widehat{\gamma}) - \sigma_1(\gamma)| = o_p(1)$.

- **Bound for** d_{k-1}. By the triangle inequality, we have

$$
\begin{aligned}
&\left|d_{k-1}(\widehat{\gamma}) - d_{k-1}(\gamma)\right| \\
&= \Big|\mathrm{prox}_{\eta_{k-1}(\widehat{\gamma})G}(\mu_{k-1}(\widehat{\gamma})\widehat{\gamma}Q_1 + \sqrt{\kappa}\sigma_{k-1}(\widehat{\gamma})Q_2 + \eta_{k-1}(\widehat{\gamma})h(\widehat{\gamma}Q_1, U)) \\
&\quad - \mathrm{prox}_{\eta_{k-1}(\gamma)G}(\mu_{k-1}(\gamma)\gamma Q_1 + \sqrt{\kappa}\sigma_{k-1}(\gamma)Q_2 + \eta_{k-1}(\gamma)h(\gamma Q_1, U))\Big| \\
&\leq \Big|\mathrm{prox}_{\eta_{k-1}(\widehat{\gamma})G}\left(\mu_{k-1}(\widehat{\gamma})\widehat{\gamma}Q_1 + \sqrt{\kappa}\sigma_{k-1}(\widehat{\gamma})Q_2 + \eta_{k-1}(\widehat{\gamma})h(\widehat{\gamma}Q_1, U)\right) \\
&\quad - \mathrm{prox}_{\eta_{k-1}(\gamma)G}\left(\mu_{k-1}(\widehat{\gamma})\widehat{\gamma}Q_1 + \sqrt{\kappa}\sigma_{k-1}(\widehat{\gamma})Q_2 + \eta_{k-1}(\widehat{\gamma})h(\widehat{\gamma}Q_1, U)\right)\Big| \\
&\quad + \Big|\mathrm{prox}_{\eta_{k-1}(\gamma)G}\left(\mu_{k-1}(\widehat{\gamma})\widehat{\gamma}Q_1 + \sqrt{\kappa}\sigma_{k-1}(\widehat{\gamma})Q_2 + \eta_{k-1}(\widehat{\gamma})h(\widehat{\gamma}Q_1, U)\right) \\
&\quad - \mathrm{prox}_{\eta_{k-1}(\gamma)G}\left(\mu_{k-1}(\gamma)\gamma Q_1 + \sqrt{\kappa}\sigma_{k-1}(\gamma)Q_2 + \eta_{k-1}(\gamma)h(\gamma Q_1, U)\right)\Big| \\
&\leq C\left|\eta_{k-1}(\widehat{\gamma}) - \eta_{k-1}(\gamma)\right| + \left|Q_1\left(\mu_{k-1}(\widehat{\gamma})\widehat{\gamma} - \mu_{k-1}(\gamma)\gamma\right)\right| \\
&\quad + \sqrt{\kappa}\left|Q_2(\sigma_{k-1}\left(\widehat{\gamma}\right) - \sigma_{k-1}(\gamma))\right| + \left|\eta_{k-1}(\widehat{\gamma})h(\widehat{\gamma}Q_1, U) - \eta_{k-1}(\gamma)h(\gamma Q_1, U)\right|,
\end{aligned}
$$

where C is some positive constant, and the last inequality follows from Lemma 10 and Lemma 9 (i). Using the triangle inequality again, we obtain

$$
\begin{aligned}
&\left|d_{k-1}(\widehat{\gamma}) - d_{k-1}(\gamma)\right| \\
&\leq \left|Q_1\mu_{k-1}(\widehat{\gamma})(\widehat{\gamma} - \gamma)\right| + \left|Q_1\left(\mu_{k-1}(\widehat{\gamma}) - \mu_{k-1}(\gamma)\right)\right|\gamma \\
&\quad + \left|\eta_{k-1}(\widehat{\gamma})Q_1 L_h\left(\widehat{\gamma} - \gamma\right)\right| + \left|\left(\eta_{k-1}(\widehat{\gamma}) - \eta_{k-1}(\gamma)\right)h(\gamma Q_1, U)\right| + o_p(1) = o_p(1).
\end{aligned}
$$

• **Bound for η_k.** Define a constant

$$C_{\eta_k} = \left(1 - \mathbb{E}\left[\frac{1}{1+\eta_{k-1}(\widehat{\gamma})g'(d_{k-1}(\widehat{\gamma}))}\right]\right)^{-1}\left(1 - \mathbb{E}\left[\frac{1}{1+\eta_{k-1}(\gamma)g'(d_{k-1}(\gamma))}\right]\right)^{-1} > 0.$$

By the triangle inequality and the technique that we use for bounding η_1, we have

$$\begin{aligned}&|\eta_k(\widehat{\gamma}) - \eta_k(\gamma)|\\ &\leq \kappa\eta_{k-1}(\gamma)C_{\eta_k}\mathbb{E}\left[|\eta_{k-1}(\widehat{\gamma})g'\left(d_{k-1}(\widehat{\gamma})\right) - \eta_{k-1}(\gamma)g'\left(d_{k-1}(\gamma)\right)|\right]\\ &\quad + O_p(|\eta_{k-1}(\widehat{\gamma}) - \eta_{k-1}(\gamma)|)\\ &\leq \kappa\eta_{k-1}^2(\gamma)C_{\eta_k}L_{g'}\mathbb{E}[O_p(|d_{k-1}(\widehat{\gamma}) - d_{k-1}(\gamma)|)] + O_p(|\eta_{k-1}(\widehat{\gamma}) - \eta_{k-1}(\gamma)|)\\ &= o_p(1).\end{aligned}$$

• **Bound for μ_k and σ_k.** By the same way to show the bound for μ_1 and σ_1, we have

$$|\mu_k(\widehat{\gamma}) - \mu_k(\gamma)| = |\sigma_k(\widehat{\gamma}) - \sigma_k(\gamma)| = o_p(1).$$

Thus, we obtain

$$|\eta_k(\widehat{\gamma}) - \eta_k(\gamma)| = |\mu_k(\widehat{\gamma}) - \mu_k(\gamma)| = |\sigma_k(\widehat{\gamma}) - \sigma_k(\gamma)| = o_p(1), \tag{22}$$

for any $k \in \mathbb{N}$ by induction.

Step 2. In this step, we get the conclusion from the results in Step 1. First, we obtain

$$|\eta(\widehat{\gamma}) - \eta(\gamma)| = |\mu(\widehat{\gamma}) - \mu(\gamma)| = |\sigma(\widehat{\gamma}) - \sigma(\gamma)| = o_p(1).$$

This follows from the fact that

$$|\eta(\widehat{\gamma}) - \eta(\gamma)| \leq |\eta(\widehat{\gamma}) - \eta_k(\widehat{\gamma})| + |\eta(\gamma) - \eta_k(\gamma)| + o_p(1),$$

by (22) and the first two terms on the right-hand side converge to zero for a large k limit. The results for μ and σ also follow in the same manner. □

Lemma 7. *Let $\widehat{\boldsymbol{\theta}}$ be the estimator, i.e. surrogate loss minimizer, in a GLM with a true coefficient vector $\boldsymbol{\theta}$ and features drawn i.i.d. from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{I}_p)$. Define (μ_n, σ_n) as in (19). Then,*

$$\frac{\widehat{\boldsymbol{\theta}} - \mu_n\boldsymbol{\theta}}{\sigma_n}$$

is uniformly distributed on the unit sphere lying in $\boldsymbol{\theta}^\perp$.

Proof of Lemma 7. Define an orthogonal projection matrix $\boldsymbol{P}_{\boldsymbol{\theta}} = \boldsymbol{\theta}\boldsymbol{\theta}^\top / \|\boldsymbol{\theta}\|^2$ onto the line including $\boldsymbol{\theta}$, and an orthogonal projection matrix $\boldsymbol{P}_{\boldsymbol{\theta}}^\perp = \boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{\theta}}$ onto the orthogonal complement of the line including $\boldsymbol{\theta}$. Let $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ be any orthogonal matrix obeying $\boldsymbol{U}\boldsymbol{\theta} = \boldsymbol{\theta}$, i.e. any rotation operator about $\boldsymbol{\theta}$. Then, since $\widehat{\boldsymbol{\theta}} = \boldsymbol{P}_{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}} + \boldsymbol{P}_{\boldsymbol{\theta}}^\perp\widehat{\boldsymbol{\theta}}$, we have

$$\boldsymbol{U}\widehat{\boldsymbol{\theta}} = \boldsymbol{U}\boldsymbol{P}_{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}} + \boldsymbol{U}\boldsymbol{P}_{\boldsymbol{\theta}}^\perp\widehat{\boldsymbol{\theta}} = \boldsymbol{P}_{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}} + \boldsymbol{U}\boldsymbol{P}_{\boldsymbol{\theta}}^\perp\widehat{\boldsymbol{\theta}}.$$

Using this, we obtain

$$\frac{\boldsymbol{U}\boldsymbol{P}_{\boldsymbol{\theta}}^\perp\widehat{\boldsymbol{\theta}}}{\|\boldsymbol{P}_{\boldsymbol{\theta}}^\perp\widehat{\boldsymbol{\theta}}\|} \overset{\mathrm{d}}{=} \frac{\boldsymbol{P}_{\boldsymbol{\theta}}^\perp\widehat{\boldsymbol{\theta}}}{\|\boldsymbol{P}_{\boldsymbol{\theta}}^\perp\widehat{\boldsymbol{\theta}}\|} = \frac{\widehat{\boldsymbol{\theta}} - \mu_n\boldsymbol{\theta}}{\sigma_n}, \tag{23}$$

where the first identity follows from the fact that $\boldsymbol{U}\widehat{\boldsymbol{\theta}} \overset{\mathrm{d}}{=} \widehat{\boldsymbol{\theta}}$ since $\boldsymbol{U}\widehat{\boldsymbol{\theta}}$ is the estimator with a true coefficient $\boldsymbol{U}\boldsymbol{\theta} = \boldsymbol{\theta}$ and features drawn i.i.d. from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. (23) reveals that $(\widehat{\boldsymbol{\theta}} - \mu_n\boldsymbol{\theta})/\sigma_n$ is rotationally invariant about $\boldsymbol{\theta}$, lies in $\boldsymbol{\theta}^\perp$, and has a unit norm. These conclude the proof. □

Lemma 8. *Suppose that* $0 < \|\boldsymbol{L}^\top\widehat{\boldsymbol{\beta}}\|, \|\boldsymbol{L}^\top\boldsymbol{\beta}\| \leq C$ *for some constant* $C > 0$. *Let* $\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}$ *be defined in* (18). *If we define*

$$\mu_n = \frac{\widehat{\boldsymbol{\theta}}^\top\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|^2}, \qquad \sigma_n^2 = \frac{1}{\kappa}\|\widehat{\boldsymbol{\theta}} - \mu_n\boldsymbol{\theta}\|^2,$$

then, under assumptions (A1) and (A2), we have

$$\mu_n \xrightarrow{\text{a.s.}} \mu, \qquad \sigma_n^2 \xrightarrow{\text{a.s.}} \sigma^2,$$

as $n, p(n) \to \infty$ *with* $p(n)/n \to \kappa$.

Proof of Lemma 8. Since $\|\boldsymbol{\theta}\|$ is bounded by an assumption, $\boldsymbol{U}\boldsymbol{\theta} = (\|\boldsymbol{\theta}\|, \dots, \|\boldsymbol{\theta}\|)/\sqrt{p}$ with some orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ satisfies an assumption of Lemma 1 with $\bar{\beta} \sim \delta_{\gamma/\sqrt{p}}$ by the fact that $\|\boldsymbol{U}\boldsymbol{\theta}\|^2 = \|\boldsymbol{\theta}\|^2 = \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} = \gamma^2$. Then applying Lemma 1 to $(\boldsymbol{U}\boldsymbol{\theta}, \boldsymbol{U}\widehat{\boldsymbol{\theta}})$ with $\psi(s,t) = st, t^2$ and considering their ratio gives

$$\frac{\langle \boldsymbol{U}\boldsymbol{\theta}, \boldsymbol{U}\widehat{\boldsymbol{\theta}}\rangle}{\|\boldsymbol{U}\boldsymbol{\theta}\|^2} = \mu_n \xrightarrow{\text{a.s.}} \mu.$$

Also, setting $\psi(s,t) = (s - \mu t)^2$ yields

$$\frac{1}{\kappa}\|\boldsymbol{U}\widehat{\boldsymbol{\theta}} - \mu\boldsymbol{U}\boldsymbol{\theta}\|^2 = \frac{1}{\kappa}\|\widehat{\boldsymbol{\theta}} - \mu\boldsymbol{\theta}\|^2 \xrightarrow{\text{a.s.}} \sigma^2.$$

Thus, we obtain

$$\sigma_n^2 = \frac{1}{\kappa}\|\widehat{\boldsymbol{\theta}} - \mu\boldsymbol{\theta}\|^2 + 2(\mu - \mu_n)\boldsymbol{\theta}^\top\widehat{\boldsymbol{\theta}} - (\mu^2 - \mu_n^2)\,\|\boldsymbol{\theta}\|^2 \xrightarrow{\text{a.s.}} \sigma^2.$$

□

Lemma 9. *Define $\|f\|_\infty := \sup_{t\in\mathbb{R}} |f(t)|$ for a function $f : \mathbb{R} \to \mathbb{R}$. A proximal operator $\text{prox}(x, f, b) \equiv \text{prox}_{bF}(x)$ with $x \in \mathbb{R}$, $b > 0$, and a monotone continuous function $f = F'$, is Lipschitz continuous with respect to*

(i) *$x \in \mathbb{R}$ with constant 1,*
(ii) *a monotone continuous function $f : \mathbb{R} \to \mathbb{R}$ in terms of L^∞ norm $\|\cdot\|_\infty$ with constant $b > 0$,*
(iii) *and $b > 0$ with constant $\|f\|_\infty$.*

Proof of Lemma 9. For any $x, y \in \mathbb{R}$ and fixed monotone $f(\cdot)$, suppose that $u = \text{prox}(x, f, b)$ and $v = \text{prox}(y, f, b)$ and $u > v$ without loss of generality. Note that $F(\cdot)$ is convex since its derivative $f(\cdot)$ is a monotonically increasing function. By the first-order condition, we have $x = u + bf(u)$ and $y = v + bf(v)$. Since $f(\cdot)$ is monotonically increasing, $0 \le (u - v)(f(u) - f(v))$. Using this,

$$0 \le (u - v)(bf(u) - bf(v)) = (u - v)(x - u - y + v) = (u - v)(x - y) - (u - v)^2,$$

by $b > 0$. This immediately implies $|u - v| \le |x - y|$, i.e., $\text{prox}(x, f, b)$ is 1-Lipschitz continuous with respect to the first argument.

Next, set any two continuous monotone functions $f(\cdot)$ and $g(\cdot)$. Suppose that $\text{prox}(x, f, b) \ge \text{prox}(x, g, b)$ with fixed $x \in \mathbb{R}$ without loss of generality. Then,

$$\begin{aligned} |\text{prox}(x, f, b) - \text{prox}(x, g, b)| &= \text{prox}(x, f, b) - \text{prox}(x, g, b) \\ &= b\left(g(\text{prox}(x, g, b)) - f(\text{prox}(x, f, b))\right) \\ &\le b\left(g(\text{prox}(x, f, b)) - f(\text{prox}(x, f, b))\right) \\ &\le b\,\|f - g\|_\infty\,, \end{aligned} \tag{24}$$

where the second identity follows from the fact that $\text{prox}(x, f) = x - bf(\text{prox}(x, f, b))$, and the first inequality is from monotinicity of $g(\cdot)$. This means $\text{prox}(x, f, b)$ is Lipschitz continuous for the second argument with constant $b > 0$ in terms of $\|\cdot\|_\infty$.

At last, using the fact that $\text{prox}(x, g, b) = \text{prox}(x, f, \alpha b)$ for $g(x) = \alpha f(x)$, $\alpha > 0$ by the definition of the proximal operator, we have, for any $b, b' > 0$,

$$\begin{aligned} |\text{prox}(x, f, b) - \text{prox}_{b'}(x, f, b')| &= \left|\text{prox}_b(x, f, b) - \text{prox}\left(x, \frac{b'}{b} f, b\right)\right| \\ &\le b \left\| f - \frac{b'}{b} f \right\|_\infty \\ &= \|bf - b'f\|_\infty \\ &= |b - b'|\,\|f\|_\infty\,, \end{aligned}$$

where the inequality follows from (24). □

Lemma 10. *Let $F : \mathbb{R} \to \mathbb{R}$ be a strictly convex function and $f = F'$. Suppose that $f(\cdot)$ takes bounded values on bounded domains. For any bounded $x \in \mathbb{R}$ and $c, c' > 0$, we have*

$$|\text{prox}_{cF}(x) - \text{prox}_{c'F}(x)| \le C\,|c - c'|\,,$$

for some positive constant C.

Proof of Lemma 10. By the proofs of Lemma 9 (ii) and (iii), we can improve Lemma 9 (iii) as

$$|\mathrm{prox}_{cF}(x) - \mathrm{prox}_{c'F}(x)| \leq |f(\mathrm{prox}_{cF}(x))|\, |c - c'|\,.$$

Thus, we can complete the proof by showing that $\mathrm{prox}_{cF}(x)$ is bounded. Remind that the definition of the proximal operator is

$$\mathrm{prox}_{cF}(x) = \operatorname*{arg\,min}_{z\in\mathbb{R}} \left\{ cF(z) + \frac{1}{2}(z-x)^2 \right\}.$$

We denote the objective function as $H(z) := cF(z)+\frac{1}{2}(z-x)^2 = H_1(z)+H_2(z)$. Obviously, H_2 has the minimum value at x.

(i) For the case that a minimizer $\tilde{z}$ of H_1 is bounded. Note that the minimizer is unique by the strict convexity of H_1. Without loss of generality, suppose that $\tilde{z} \leq z$. In this case, we have $H_1'(z') < 0$ and $H_2'(z') < 0$ for $z' < \tilde{z}$, and also have $H_1'(z') > 0$ and $H_2'(z') > 0$ for $z < z'$. Hence, $H'(z') < 0$ holds for $z' < \tilde{z}$ and $H'(z') > 0$ holds for $z < z'$, thus $\mathrm{prox}_{cF}(x) \notin [-\infty, \tilde{z}) \cup (z, \infty]$. In contrast, $H'(z')$ may be both positive or negative for $z' \in [\tilde{z}, z]$. Hence, $\mathrm{prox}_{cF}(x) \in [\tilde{z}, z]$ holds and thus bounded itself.

(ii) For the case when H_1 has a minimum at an unbounded point, such as $H_1(z) = ce^z$, we can also show that $\mathrm{prox}_{cF}(x)$ is bounded. In this case, we can assume that H_1 is monotonically increasing without loss of generality. For $z > x$, we have $H'(z) > 0$ by $H_1'(z) > 0$ and $H_2(z) > 0$. Since $H_1'(z)$ is positive and monotonically increasing by the monotonicity and the convexity of $H_1(z)$, there exists a constant $\tilde{C} < x$ such that $H_1'(z) = cf(z) < -H_2'(z) = -(z-x)$ for any $z < \tilde{C}$. Putting together the results, we have $\mathrm{prox}_{cF}(x) \in [\tilde{C}, x]$. □

References

[AHW95] Peter Auer, Mark Herbster, and Manfred KK Warmuth. Exponentially many local minima for single neurons. *Advances in neural information processing systems*, 8, 1995.

[AKK+14] Alekh Agarwal, Sham Kakade, Nikos Karampatziakis, Le Song, and Gregory Valiant. Least squares revisited: Scalable approaches for multi-class prediction. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2014.

[Bel22] Pierre C Bellec. Observable adjustments in single-index models for regularized m-estimators. *arXiv preprint arXiv:2204.06990*, 2022.

[BK25] Pierre C Bellec and Takuya Koriyama. Error estimation and adaptive tuning for unregularized robust m-estimator. *Journal of Machine Learning Research*, 26(16):1–40, 2025.

[BKM+19] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[BM11] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[Bol14] Erwin Bolthausen. An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[BS22] Pierre C Bellec and Yiwei Shen. Derivatives and residual distribution of regularized m-estimators with application to adaptive tuning. In *Conference on Learning Theory*, pages 1912–1947. PMLR, 2022.

[BZ21] Pierre C Bellec and Cun-Hui Zhang. Second-order stein: Sure for sure and other applications in high-dimensional inference. *The Annals of Statistics*, 49(4):1864–1903, 2021.

[CGM21] T Tony Cai, Zijian Guo, and Rong Ma. Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association*, pages 1–14, 2021.

[CGM23] T Tony Cai, Zijian Guo, and Rong Ma. Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association*, 118(542):1319–1332, 2023.

[Cha09] Sourav Chatterjee. Fluctuations of eigenvalues and second order poincaré inequalities. *Probability Theory and Related Fields*, 143(1-2):1–40, 2009.

[CLM24] Xingyu Chen, Lin Liu, and Rajarshi Mukherjee. Method-of-moments inference for glms and doubly robust functionals under proportional asymptotics. *arXiv preprint arXiv:2408.06103*, 2024.

[Cor83] Gauss M Cordeiro. Improved likelihood ratio statistics for generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(3):404–413, 1983.

[CS20] Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42, 2020.

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[DJS$^+$89] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5):304–310, 1989.

[DM16] David Donoho and Andrea Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969, 2016.

[DMM09] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the*

National Academy of Sciences, 106(45):18914–18919, 2009.

[EK18] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170:95–175, 2018.

[EKBB$^+$13] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[FL21] Zhe Fei and Yi Li. Estimation and inference for high dimensional generalized linear models: A splitting and smoothing approach. *Journal of Machine Learning Research*, 22(58):1–32, 2021.

[FVRS22] Oliver Y Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J Samworth. A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning*, 15(4):335–536, 2022.

[GC16] Bin Guo and Song Xi Chen. Tests for high dimensional generalized linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 1079–1102, 2016.

[HMRT22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[Jør87] Bent Jørgensen. Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(2):127–145, 1987.

[JVDG16] Jana Janková and Sara Van De Geer. Confidence regions for high-dimensional generalized linear models under sparsity. *arXiv preprint arXiv:1610.01353*, 2016.

[KR58] Leonid Vasilevich Kantorovich and SG Rubinshtein. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59, 1958.

[LGC$^+$21] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

[LZCL23] Sai Li, Linjun Zhang, T Tony Cai, and Hongzhe Li. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, pages 1–12, 2023.

[McC80] Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

[ML19] Xiaoyi Mai and Zhenyu Liao. High dimensional classification via regularized and unregularized empirical risk minimization: Precise error and optimal loss. *arXiv preprint arXiv:1905.13742*, 2019.

[MLC19] Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361. IEEE, 2019.

[MM21] Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335, 2021.

[MM22] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[MS22] Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

[PSVAV24] Jolien Ponnet, Pieter Segaert, Stefan Van Aelst, and Tim Verdonck. Robust inference and modeling of mean and dispersion for generalized linear models. *Journal of the American Statistical Association*, 119(545):678–689, 2024.

[Ran11] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE, 2011.

[RSR$^+$16] Sundeep Rangan, Philip Schniter, Erwin Riegler, Alyson K Fletcher, and Volkan Cevher. Fixed points of generalized approximate message passing with arbitrary matrices. *IEEE Transactions on Information Theory*, 62(12):7464–7474, 2016.

[SAH19] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[SC19] Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[SCC19] Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability theory and related fields*, 175(1):487–558, 2019.

[SLCT21] Mohamed El Amine Seddik, Cosme Louart, Romain Couillet, and Mohamed Tamaazousti. The unexpected deterministic and universal behavior of large softmax classifiers. In *International Conference on Artificial Intelligence and Statistics*, pages 1045–1053. PMLR, 2021.

[SUI24] Kazuma Sawaya, Yoshimasa Uematsu, and Masaaki Imaizumi. High-dimensional single-index models: Link estimation and marginal inference. *arXiv preprint arXiv:2404.17812*, 2024.

[TAH15] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. *Advances in Neural Information Processing Systems*, 28, 2015.

[TAH18] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[TB23] Kai Tan and Pierre C Bellec. Multinomial logistic regression: Asymptotic normality on null covariates in high-dimensions. *arXiv preprint arXiv:2305.17825*, 2023.

[TK18] Takashi Takahashi and Yoshiyuki Kabashima. A statistical mechanics approach to de-biasing and uncertainty estimation in lasso for random measurements. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(7):073405, 2018.

[TT17] Xiaoying Tian and Jonathan Taylor. Asymptotics of selective inference. *Scandinavian Journal of Statistics*, 44(2):480–499, 2017.

[VdGBRD14] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, pages 1166–1202, 2014.

[vdV00] Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

[VKM22] Ramji Venkataramanan, Kevin Kögler, and Marco Mondelli. Estimation in rotationally invariant generalized linear models via approximate message passing. In *International Conference on Machine Learning*, pages 22120–22144. PMLR, 2022.

[YYMD21] Steve Yadlowsky, Taedong Yun, Cory Y McLean, and Alexander D'Amour. Sloe: A faster method for statistical inference in high-dimensional logistic regression. *Advances in Neural Information Processing Systems*, 34:29517–29528, 2021.

[ZSC22] Qian Zhao, Pragya Sur, and Emmanuel J Candes. The asymptotic distribution of the mle in high-dimensional logistic models: Arbitrary covariance. *Bernoulli*, 28(3):1835–1861, 2022.